PB-240 214

PROCEEDINGS OF THE SYMPOSIUM ON STATISTICAL ASPECTS OF AIR QUALITY DATA, HELD AT THE CAROLINA INN, CHAPEL HILL, NORTH CAROLINA ON NOVEMBER 9-10, 1972

Lawrence D. Kornreich

Triangle Universities Consortium on Air Pollution

Prepared for:

Air Pollution Technical Information Center

October 1974

DISTRIBUTED BY:

NTIS

National Technical Information Service
U. S. DEPARTMENT OF COMMERCE

Research reports of the Office of Research and Development, Environmental Protection Agency, have been grouped into five series. These five broad categories were established to facilitate further development and application of environmental technology. Elimination of traditional grouping was consciously planned to foster technology transfer and a maximum interface in related fields. The five series are:

1. Environmental Health Effects Research
2. Environmental Protection Technology
3. Ecological Research
4. Environmental Monitoring
5. Socioeconomic Environmental Studies

This report has been assigned to the ENVIRONMENTAL MONITORING series. This series describes research conducted to develop new or improved methods and instrumentation for the identification and quantification of environmental pollutants at the lowest conceivably significant concentrations. It also includes studies to determine the ambient concentrations of pollutants in the environment and/or the variance of pollutants as a function of time or meteorological factors.

Copies of this report are available free of charge to Federal employees, current contractors and grantees, and nonprofit organizations - as supplies permit - from the Air Pollution Technical Information Center, Environmental Protection Agency, Research Triangle Park, North Carolina 27711. This document is also available to the public for sale through the Superintendent of Documents, U. S. Government Printing Office, Washington, D. C. 20402.

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before comp

PB 240 214

1. REPORT NO. EPA-650/4-74-038
2.
4. TITLE AND SUBTITLE
Proceedings of the Symposium on Statistical Aspects of Air Quality Data
5. REPORT DATE
October 1974
6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S)
Lawrence D. Kornreich, Editor, Executive Director, Triangle Universities Consortium on Air Pollution
8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS
Triangle Universities Consortium on Air Pollution
Post Office Box 2284
Chapel Hill, North Carolina 27514
10. PROGRAM ELEMENT NO.
1AA009
11. CONTRACT/GRANT NO.
68-02-0994
12. SPONSORING AGENCY NAME AND ADDRESS
Meteorology Laboratory
National Environmental Research Center
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711
13. TYPE OF REPORT AND PERIOD COVERED
Symposium Proceedings
14. SPONSORING AGENCY CODE
15. SUPPLEMENTARY NOTES
APTIC # T5067
16. ABSTRACT

The 15 papers in these proceedings analyze air quality data as a function of frequency, maxima, the form of the frequency distribution, and averaging time. Concentrations and frequency distributions calculated with meteorologic diffusion models are compared with observed values. Discussions that followed the paper presentations are included.

Reproduced by
NATIONAL TECHNICAL INFORMATION SERVICE
US Department of Commerce
Springfield, VA. 22151

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS		b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Air quality data Statistical analyses Frequency distribution Lognormal Averaging time Meteorology	Diffusion Sulfur dioxide Suspended particulate		

18. DISTRIBUTION STATEMENT Unlimited	19. SECURITY CLASS *(This Report)* Unclassified	21. NO. OF PAGES 266
	20. SECURITY CLASS *(This page)* Unclassified	22. PRICE

EPA Form 2220-1 (9-73)

☆ G.P.O. 1973 — 640-878 / 659. REGION NO. 4

EPA-650/4-74-038

PROCEEDINGS
OF THE SYMPOSIUM
ON STATISTICAL ASPECTS
OF AIR QUALITY DATA

Editor:
Dr. Lawrence D. Kornreich
Executive Director, Triangle Universities
Consortium on Air Pollution

Symposium Sponsors:
Meteorology Laboratory, National Environmental Research Center
and
Triangle Universities Consortium on Air Pollution

Contract No. 68-02-0994
(with University of North Carolina, Chapel Hill)
ROAP No. 21ADO
Program Element No. 1AA009

EPA Project Officer: Dr. Ralph Larsen

Prepared for

U.S. ENVIRONMENTAL PROTECTION AGENCY
Office of Research and Development
National Environmental Research Center
Research Triangle Park, N.C. 27711

October 1974

ia

This report has been reviewed by the Environmental Protection Agency and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

Publication No. EPA-650/4-74-038

PREFACE

The symposium on Statistical Aspects of Air Quality was held on November 9 and 10, 1972, at the Carolina Inn, in Chapel Hill, N. C., in accordance with the terms of a contract between the Division of Meteorology* of the U. S. Environmental Protection Agency (EPA) and the University of North Carolina at Chapel Hill (UNC).

Although UNC was the contractor, it was agreed that the symposium would be sponsored by the Triangle Universities Consortium on Air Pollution (TUCAP), an association of Duke University, North Carolina State University and the University of North Carolina at Chapel Hill. The project officer for EPA was Mr. Charles R. Hosler; the responsible officer for TUCAP was Dr. Lawrence D. Kornreich. The detailed planning for the symposium was done by a steering committee representing both TUCAP and EPA.

All papers that were presented at the symposium are included in this volume. Most of the technical papers were reprinted and distributed to the participants prior to the meeting. An open discussion followed the presentation of each technical paper, and the questions and answers were recorded and transcribed. Each discussant was given the opportunity to review and edit his comments. Only those comments which were reviewed and approved by the discussants appear in this volume.

For their outstanding performances at the banquet, special thanks are due Mr. Donald Pack of the National Oceanographic and Atmospheric Administration, who acted as Toastmaster, and Dr. Roy Kuebler of the UNC Department of Biostatistics who was the featured speaker.

The registration of participants and preparation of information packets was ably handled by Continuing Education and Field Service of the UNC School of Public Health. The audio-visual arrangements throughout the meeting were capably managed by Mr. Lewis Kontnick and Mr. E. James Dale, graduate students in the air pollution curriculum at UNC.

I am especially grateful to Professor Arthur C. Stern of the UNC Department of Environmental Sciences for his help and guidance in preparing this volume. I also want to thank my secretaries for their outstanding service—Mrs. Jean Lang during the planning and holding of the symposium, and Mrs. Ann Harrell during the preparation of this volume.

Lawrence D. Kornreich
Chapel Hill, N. C.
February, 1974

---

*Now the Meteorology Laboratory of the National Environmental Research Center, Research Triangle Park, N. C.

STEERING COMMITTEE

Kenneth L. Calder, Chief Scientist, Division of Meteorology, Environmental Protection Agency.

Kenneth R. Knoerr, Associate Professor, Biometeorology, Duke University.

Lawrence D. Kornreich, Executive Director, Triangle Universities Consortium on Air Pollution

Ralph I. Larsen, Environmental Research Engineer, Division of Meteorology, Environmental Protection Agency.

Arthur C. Stern, Professor of Air Hygiene, University of North Carolina at Chapel Hill

Allen H. Weber, Associate Professor, Geosciences, North Carolina State University

SESSION CHAIRMEN

Irving Singer, Smith-Singer Meteorologists

Donald Rote, Argonne National Laboratory

James Arvesen, Purdue University

Arnold Court, California State University

George Tiao, University of Wisconsin

Ray Wanta, Consulting Meteorologist

WORKSHOP LEADERS

Ralph Larsen, Division of Meteorology, U. S. Environmental Protection Agency

Donald McNeil, Princeton University

Warren Johnson, Division of Meteorology, U. S. Environmental Protection Agency

Victor Hasselblad, Human Studies Laboratory, U. S. Environmental Protection Agency

CONTENTS

1. *Robert A. McCormick* 1-1
Welcoming Remarks

2. *Arthur C. Stern* 2-1
Keynote Address: Statistical Analysis of Air Quality Data

3. *Frank A. Gifford, Jr.* 3-1
The Form of the Frequency Distribution of Air Pollution Concentrations

4. *F. K. Wipperman* 4-1
Meteorological Parameters Relevant in a Statistical Analysis of Air Quality Data

5. *Michael Benarie* 5-1
The Use of the Relationship between Wind Velocity and Ambient Pollutant Concentration Distributions for the Estimation of Average Concentrations from Gross Meteorological Data

6. *Richard E. Barlow and Nozer D. Singpurwalla* 6-1
Averaging Time and Maxima for Dependent Observations

7. *Allan H. Marcus* 7-1
A Stochastic Model for Estimating Pollutant Exposure by Means of Air Quality Data

8. *Harold E. Neustadter, Steven M. Sidik and John C. Burr, Jr.* 8-1
Evaluating Conformity with Two-Point Air Quality Standards, Polludex

9. *Joseph B. Knox and Richard I. Pollack* 9-1
An Investigation of the Frequency Distributions of Surface Air-Pollutant Concentrations

10. *D. Bruce Turner* 10-1
Air Quality Frequency Distributions from Dispersion Models Compared with Measurements

11. *Bernard E. Saltzman* 11-1
Fourier Analysis of Air Monitoring Data

12. *F. Barry Smith and G. H. Jeffrey* 12-1
The Prediction of High Concentrations of Sulfur Dioxide in London and Manchester Air

13. *David A. Lynn* 13-1
Fitting Curves to Urban Suspended Particulate Data

14. *Joseph R. Visalli, David L. Brenchley and Howard Reiquam* 14-1
A Proposed Ambient Air Quality Sampling Strategy and Methodology for the Design of Surveillance Networks

15. *Yuji Horie and John H. Overton* 15-1
The Effect on Rollback Models Due to Distribution of Pollutant Concentrations

16. Symposium Participants 16-1

17. Bibliographic Data Sheet 17-1

1. WELCOMING REMARKS

ROBERT A. McCORMICK*

Division of Meteorology
National Environmental Research Center
Environmental Protection Agency
Research Triangle Park, North Carolina

On behalf of the National Environmental Research Center in the Research Triangle Park, North Carolina, and in particular the Division of Meteorology who have actually sponsored the Symposium, I would like to warmly welcome you all and say how pleased we are that so many distinguished investigators have found time to attend, especially those from overseas.

Before Professor Stern's opening remarks, I would like to say a few words as to why we in the Division of Meteorology (DMT)** were anxious to support this Symposium. Following the pioneering efforts of Fran Pooler and Bruce Turner, our primary efforts have been in the development of source-oriented diffusion-type models. Because of their source-oriented structure, they have potentially wide application in the consideration of the effects on air quality of hypothetical and arbitrary emission control strategies. They thus provide a rational basis for air quality management through control of selected sources of pollution. At the present time the development and improvement of such air quality simulation models has the highest priority of all items in the research program of the DMT.

In contrast are those models that primarily involve some form of statistical regression analysis and that depend entirely on the availability of extensive meteorological and air quality data for a particular urban location. Over the years these statistical approaches have become increasingly sophisticated and now include such things as "multiple-discriminant analysis", "empirical orthogonal functions", "factor analysis", and most recently "computerized adaptive pattern classification". Although these developments have had useful applications to specific problems, the fact that they are receptor- rather than source-oriented and do not involve any explicit input of information concerning pollution emissions, *so far* makes them *not* applicable to comparative studies of control strategies. This is the reason for their lower priority in our research program as compared with the source-oriented dispersion-type models. I have intentionally said "so far" in the preceding as it seems to us that a question which ought to be resolved is whether some form of statistical predictive scheme could be evolved that would incorporate hypothetical changes in the pollution emissions distribution without any direct use of meterological diffusion theory. Such a model would, of course, then possess the desirable source-oriented property.

---

*On assignment from the National Oceanic and Atmospheric Administration, U. S. Department of Commerce at the time of the Symposium. Mr. McCormick has since retired from Federal service.

**Now the Meteorology Laboratory.

By analogy with the striking developments of the last decade or so in turbulent fluid mechanics, we can hope that advances and improved understanding of air quality simulation will result from a marriage of statistical techniques with more precise physical formulations of the problems. It was this strong feeling that suggested the need for the present Symposium. Perhaps this will stimulate a stronger interaction between some of the more meteorologically inclined of our air quality modelers and non-meteorological statistical experts. No less important is the hope that the workshop sessions will more clearly define for our air quality modeling fraternity, those areas and approaches where improved cooperation between meteorologist and statistician might be most helpful and fruitful in the immediate future, in much the same manner as the clarification now being achieved between the meteorologists and atmospheric chemists.

Keynote Address

2. STATISTICAL ANALYSIS OF AIR QUALITY DATA

ARTHUR C. STERN

Department of Environmental Sciences and Engineering
University of North Carolina
Chapel Hill, North Carolina

Introduction

The need for a Symposium on Statistical Analysis of Air Quality Data arises from a quite diverse assortment of challenges to our understanding of the meaning of air quality data now coming into sharper focus. A closer look at these challenges will give a better understanding of some of the answers we are seeking at this Symposium.

Much of the research effort in the field of air pollution is and has been directed to the development of air quality criteria, and through them to the establishment of air quality standards. Since air quality criteria are explorations of the relationship between levels of air quality and the adverse effects found in receptors exposed to these levels, it is essential that there be precise description of both these levels and their associated adverse effects. It is not the function of this Symposium to discuss the precise description of the adverse effects, but it is our function to discuss the precise description of the levels of exposure that cause them.

Simple Chamber Atmospheres

Our simplest task is to describe the exposure in a chamber in which one or more receptors are being experimentally exposed. These receptors may variously be materials specimens, such as textiles, paper or leather; vegetation, such as plants, lichens, or bacteria; animals, such as monkeys, guinea pigs, or mice; or human volunteers. However, even in this most simple situation an adequate description of the quality of the air in the chamber must reflect the variance in the system generating the chamber atmosphere; the decay in contaminant level in the chamber due to wall effects and absorption by chambers contents, such as the fur of exposed animals, their urine and feces, and cyclic response associated with the effect of activity on animal uptake and light-intensity on plant uptake.

Complex Chamber Atmospheres

Description of the level of exposure becomes more complicated when long-term exposure is adjusted to the working hours of the experimenters, as when exposures are for 8 hours a day, 5 days a week, with occasional hours or days lost due to staff holidays, equipment maintenance or malfunction, or other causes.

A still higher level of skill is needed to describe the level that results when the experimenter, instead of attempting to maintain a constant level, deliberately attempts to simulate in the chamber some of the changes in air quality level observed in the ambient air. This is a problem with which some of my colleagues at UNC-Chapel Hill are now having to cope in connection with the introduction of reactants into a 12,000 cubic foot, naturally irradiated "Teflon" chamber under construction a few miles from here. The intent is to introduce the reactants at a rate that will simulate what happens in the ambient air of a community such as Los Angeles on a typical weekday morning. Precise description of the quality of the contents of the chamber will be as complex as describing the quality of ambient air.

In effects research, be it of materials, plants, animals, or humans, one of the things we most need to know is the relative influence on the adverse effects observed of short-duration, high-concentration spikes superimposed on long sustained average levels. Very few, if any, chamber experiments attempt this type of superimposition and raise the problem of air quality description which it would impose. The importance of this kind of information arises because we need to know in which classes of receptors adverse effects are proportional to integrated dose, and in which classes protective or defense mechanisms are inhibited by short-duration, high-concentration spikes so that adverse effect is more than proportional to integrated dose.

Ambient Air Quality Data for Specific Effects

As great as is the challenge of providing precise description of exposure chamber atmospheres, even more challenging is the task of precisely describing the ambient air. Air quality criteria development also requires that the exposure of materials, livestock, forests, crops, and human populations to the ambient air be described during both short-term episodic conditions and for long periods of time, up to the lifetimes of viable receptors. Here we have three situations of increasing complexity. The first, and least complex, is providing a description of air quality at a fixed location in the field where material specimens are being exposed; test crops are being grown; or animals are being maintained under observation and air quality is being monitored.

Next in level of complexity is providing a description of the quality of the air associated with a specific observed adverse effect occurring at other than a fixed experimental location, as was the former case. Typical situations are:

(a) Gradual or sudden increase in clinic or hospital admissions for asthmatic attack, respiratory or cardiac disease.

(b) Gradual or sudden awareness of specific damage to trees, crops, or livestock.

(c) Gradual or sudden awareness of specific damage to materials.

In such situations, it is rare that air was being monitored at the precise location where the person who was hospitalized lived, the damaged crop was being grown, or the damaged material was in use. However, part of the reason why we are here today is to better define the kinds of air quality measurements that must be made to provide air quality data to relate to these types of situations. A closely related reason for our Symposium is the as-yet-unresolved role of air quality measurement in the prediction and management of short-term air pollution episodes. There is no question but that we need good description of air quality to understand what has happened and is happening during air episodes. However, the extent that air quality data can be used to forecast the occurrence or course of an episode is still not clear and is an area into which a group such as this should be able to provide some insight.

Ambient Air Quality Data for Epidemiological Studies

Our most difficult and imperative task in the statistical analysis of air quality data is to provide descriptors of air quality that are meaningful for understanding epidemiological data on human mortality and morbidity, since this is the description of air safe to breathe. We need to know where to measure, what to measure, how frequently to measure, and how to analyze and interpret the data we measure.

Once air quality criteria are endowed with meaningful descriptions of air quality, we are in a position to select some of these descriptors as air quality standards. To date, we have chosen such descriptors rather sparingly, using only averaging time and simple statements of frequency of occurrence. Can you supply better ones?

Federal Air Quality Standards for CO, HC, and NO_2

So much for the generalities. Now let's get down to a few examples of specifics—to some real problems created by the way air quality data and air quality standards are presently described. Among these problems have been the descriptors used for the Federal Air Quality Standards for CO, HC and NO_2. First let's look at the CO standard of 9 ppm (8-hour average), not to be exceeded more than once a year. This is ambiguous since it does not specify which 8-hour period to use. There are an infinite number of possible 8-hour running average values if the running averages may start randomly, not necessarily on the hour, at any instant of time during the year. In practice, to determine compliance with this national standard, data must be organized. Since there is also a 1-hour average national standard, the unit of time apparently intended was the clock hour rather than any randomly started period of 60 minutes. For clock hour data, the 8-hour average possibilities start with the use of one specific 8-hour period (e.g. 8 a.m. - 4 p.m.) per calendar day. Since there are 365 such periods possible per year, if one period exceeds the standard, the standard represents the 99.7 percentile value. There are 1095 possible values. There are 1095 possible non-overlapping 8-hour periods per calendar year, and between 8752 and 8760 possible running 8-hour periods, which if similarly used would set the standard at the 99.9 and 99.99 percentiles, respectively. It is thus unclear whether the 99.7, 99.9 or 99.99 percentile values were intended as the standard. The problem for the person who must establish compliance with the National standard is then to determine which of these possibilities was intended and is acceptable.

Next let's look at the NO_2 standard of 0.05 ppm—annual average. In order to relate this value meaningfully to the National standard for hydrocarbons (non-methane) of 0.24 ppm (3-hour average—6-9 a.m.), and to the NO_x reduction required in automobiles, it is necessary to convert the Federal NO_2 standard to its equivalent 3-hour average NO_x value. This requires a double conversion—first of annual average NO_2 to equivalent 3-hour average NO_2 then to equivalent 3-hour average NO_x. The hydrocarbon standard with which this latter value must be considered is for non-methane hydrocarbons, while the National automobile emission standard which is intended to achieve it, is for hydrocarbons including methane. Finally the National oxidant standard that is intended to be achieved by the control of HC (3-hour average) and NO_2 (annual average) is expressed as a 1-hour average. We hope that this Symposium will help provide more rational bases for understanding and expressing air quality data and air quality standards.

Conclusion

In October, 1969, we ran a predecessor to this Symposium, one on Multiple-Source Urban Diffusion Models. There are several important ties between these two symposia. Both were under the same joint sponsorship of the Division of Meteorology (EPA) and the Triangle Universities Consortium on Air Pollution. Some of the areas opened up at the 1969 symposium form the basis for research papers at this one. Finally as diffusion modeling comes of age, its requirements for real air quality data for model calibration and validation become increasingly important. Some of the air quality data analysis techniques discussed here should help improve the quality of diffusion models just as diffusion modeling should help improve the analysis, tabulation and presentation of air quality data.

One can envision a combined air quality monitoring-air quality modeling effort in which more accurate air quality data for a community can be obtained at lower cost in monitoring equipment and manpower, by using the model to fill in the data where there are no monitoring stations, and using the monitoring stations to calibrate and validate the model. In the past, air quality monitoring and modeling have been considered quite separate and disparate operations. What is proposed is that they can be operated as a joint enterprise with benefit to both aspects.

If this Symposium can point the way to these kinds of interactions, we may well be planting the seed from which another symposium on such interactions may arise a few years hence, just as this one arose from the seeds planted in 1969.

3. THE FORM OF THE FREQUENCY DISTRIBUTION OF AIR POLLUTION CONCENTRATIONS

FRANK A. GIFFORD

Air Resources Atmospheric Turbulence & Diffusion Laboratory
National Oceanic & Atmospheric Administration
Oak Ridge, Tennessee

Introduction

The practical need to be able to estimate the frequency distribution of air pollution concentrations doesn't have to be pointed out to the participants in this symposium. It will I'm sure be emphasized in many of the papers that we will hear. Instead, I'm going to try to bring out some of the possible reasons for the concentration frequency distributions that we observe. Along the way I hope to emphasize the basic difference between the frequency distribution of urban air pollution, which results from the combined effects of many sources, and that of concentrations from a single, isolated source of pollution, like an electric power plant. I'll also mention several different proposals that have been given in the literature for the mathematical form of these distributions, and will try to bring out some relationships among several of these.

The Lognormal Distribution of Urban Air Pollution

Larsen (1970; 1971) has established, by means of a large number of data comparisons, that observed air pollution concentration distributions are closely approximated by the lognormal function. This is an interesting fact about air pollution, which calls for some kind of explanation.

The concentration X due to an urban air pollutant, the ambient air quality in other words, is ordinarily observed over successive, short, time intervals. A record of X at an air pollution sampling point consists of a series of observed values, X_i, i = 0,1,2, . . . n. The irregular change in these numbers from one sampling interval to the next reflects all the obviously complex variability of source, meteorological, and other factors. But at any time it will most strongly depend on the existing air pollution concentration level. This is true for at least two reasons. First, relevant meteorological factors, principally the wind speed and direction, tend to be strongly self-correlated. Second, urban air pollution is more uniform than that from isolated rural sources because the urban source is distributed more uniformly, over a very large area.

This suggests that the X_i are generated by the following simple process:

$$X_1 = X_0 + y_1 X_0; \quad X_2 = X_1 + y_2 X_1; \quad \ldots, \quad X_n = X_{n-1} + y_n X_{n-1} \tag{1}$$

The quantities y_i are specified to be irregular, stochastic variables, known only in terms of their means and standard deviations. Nothing more definite than this needs to be said about the y_i, but the implication is that y_i results from a large number of small, irregular effects acting on X_{i-1}, such as brief shifts in the wind, changes in traffic, and so on. The lognormal distribution of X follows directly.

From Equation 1 it is seen by rearranging that

$$\sum_{i=1}^{n} y_i \equiv z_r = \sum_{i=1}^{n} (X_i - X_{i-1}) / X_{i-1} \tag{2}$$

According to the central limit theorem, z_n, the sum of n independent stochastic variates, will be normally distributed for large n; but Equation 2 is equivalent to

$$z_n = \int_{X_0}^{X_n} X^{-1} \, dX = \ln (X_n / X_0) \tag{3}$$

That is, X_n has a lognormal distribution. The conditions of validity of the central limit theorem are very general. The y_i do not have to be normally distributed, or even to have the same distribution.

This derivation of the lognormal distribution is just a particularization of the standard explanation of how skew distributions are generated; see for example Hald (1952). Since the cause of the irregular changes in y_i did not have to be specified exactly, the derivation only gives the form and not the parameters of the distribution. Nevertheless it seems to be an adequate explanation of the observed strong tendency towards lognormality of air pollution concentration distributions.

The Distribution of Concentration from an Isolated Point Source of Pollution

In some contrast to the above simple, but rather general result, I proposed a specific model of the frequency distribution of concentrations from an isolated point source (Gifford (1959)). See also Scriven (1965). In this model, the concentration X_p at a point (x, y, z) due to a fluctuating plume of contaminant originating at (0, 0, 0) is given by a randomly positioned, spreading-disk, Gaussian plume model:

$$X_p / Q = (2\pi\sigma_y\sigma_z U)^{-1} \exp -\left[(y - D_y)^2 / 2\sigma_y^2 + (z - D_z)^2 / 2\sigma_z^2\right] \tag{4}$$

Q is the source strength, U is the (constant) mean wind speed, and D_y and D_z are the fluctuating distances of the center of the instantaneous plume from the mean plume axis (x, 0, 0). D_y and D_z are assumed to be normally distributed, and σ_y and σ_z are the standard deviations of the instantaneous plume spreading. If the new variables

$$Y = (y - D_y)/(2\sigma_y^2)^{1/2}, \text{ and } Z = (z - D_z)/(2\sigma_z^2)^{1/2},$$

are defined, it follows from Equation 4 that

$$L = Y^2 + Z^2 = -\ln(cX_p/Q) \tag{5}$$

where

$$c = 2\pi \sigma_y \sigma_z U$$

For suitably standardized values of the variables, and for concentrations measured on or near the mean plume axis, it was demonstrated that the distribution of L is

$$p(L) = e^{-L/2}/2 \tag{6}$$

Thus the logarithm of concentration from a single source is distributed exponentially, which is the same as chi-square with two degrees of freedom. The degrees of freedom correspond to the two directions, y and z, into which the plume fluctuations are resolved by the model. The mathematical form of the distribution for sampling points off the mean axis turns out to be considerably more complicated. However the parameters of the distribution, the mean and standard deviation, are given explicity in terms of the plume parameters. See the references for details.

Gartrell (1966) pointed out that the distribution of concentrations from an isolated source is qualitatively different from that due to urban air pollution. The most probable concentration from an isolated source is clearly zero, and the distribution is strongly skewed. Gartrell found that a semilogarithmic diagram, in which concentration is plotted as the logarithm of frequency, yielded a good linear correlation of large amounts of TVA SO_2, i.e., essentially isolated point-source, data. Urban air pollution distributions are, on the other hand, flatter, and richer in the low concentration range with a higher modal value; high concentrations are less frequent.

More recently, Barry (1971) has also plotted extensive amounts of argon-41 data semilogarithmically and shows an excellent fit to the semilogarithmic distribution

$$P(X_p) = a\,e^{bX_p} \tag{7}$$

The sources of these concentrations are isolated, tall stacks. Since there is little or no background contamination, these give the isolated point source in a particularly clear-cut example. He and Scriven (1971) discussed this result, concluding that because of the simplicity of Equation 7 and the high quality of the agreement with data, this empirical, semilogarithmic distribution is to be preferred to Equation 6. Actually the two distributions refer to different quantities.

The mean wind is assumed to be constant in Equation 6, over a time period of the order of an hour, whereas in Equation 7 all wind fluctuations are included. For this reason Equation 7 is a more practically useful result. There must be some fairly direct relationship between the two results, but so far it has not been found.

The Distribution from n Point Sources

If there are a large number, n, of sources of a pollutant whose plumes affect a particular point, the i-th one will make a contribution to the concentration at the point given by

$$X_{pi}/Q_i = (2\pi \sigma_{iy} \sigma_{iz} U)^{-1} \exp\left\{-\left[(y_i - D_{yi})^2 / 2\sigma_{yi}{}^2 + (z_i - D_{zi})^2 / 2\sigma_{zi}\right]\right\} \tag{8}$$

from Equation 4. Following the same procedure,

$$-\ln (C_i X_{pi}/Q_i) = Y_i^2 + Z_i^2 \tag{9}$$

Summing over n such sources gives

$$-L = \ln \prod_{i=1}^{n} c_i X_{pi}/Q_i = \sum_{i-1}^{p} (Y_i^2 + Z_i^2) \tag{10}$$

If the quantities D_{yi}, D_{zi} are independently normally distributed with mean = 0 and $\sigma = (D^2)^{1/2}$, then the quantities $(y_i - D_{yi})^2$ and $(z_i - D_{zi})^2$ are also normally distributed, with mean = $y_i/(2\sigma_{yi}{}^2)^{1/2}$ and $\sigma = (D_{yi}^2/2\sigma_{yi}{}^2)^{1/2}$ and similarly for the z-term. By the central limit theorem, the quantity L is therefore asymptotically normally distributed, with mean and standard deviation obtained by summing those for the individual sources. Equation 10 can be written

$$\ln \left[\prod_{i=1}^{n} c_i X_{pi}/Q_i\right]^{1/n} = -L/n \tag{11}$$

and the quantity in parentheses is just the geometric mean of the concentration, weighted by c_i. If the arithmetic mean is simply related to the geometric mean (for instance if they are proportional), then Equation 11 says that the logarithm of the concentration due to a large number of point sources is normally distributed.

A distribution function for point sources based on the Poisson distribution was proposed by Wipperman (1966). His model essentially assumed a uniform, "top-hat" distribution of the instantaneous plume, which lends itself very neatly to the Poisson representation. Similarly Prinz and Stratmann (1966) developed a model using the negative binomial distribution. These have also been used to describe multiple-source, urban pollution data. Probably any of these general, skewed, frequency distributions could be used successfully to describe urban air pollution distributions. Most of the empirical comparisons, as a result of

Larsen's extensive studies, have been made with the lognormal distribution. The fit of urban pollution data to the lognormal curve, while good, is not perfect. Systematic departures occur, particularly for low concentration values. For this reason extrapolation of the lognormal, or any other distribution function out of the usual range of observed frequencies, should be made cautiously.

A final point is that many observed air pollution frequency distributions must be composites, reflecting the presence of both the multiple, distributed, urban pollution sources and nearby, strong, isolated point sources as well, in varying degrees. Study of all the resulting distribution types should be rewarding, not only for their theoretical interest but also as clues to the nature of urban air pollution sources. Concentration distributions are, so to speak, the "fingerprints" of air pollution, and their characteristics may help us detect and analyze urban air pollution patterns.

Acknowledgement

This research was performed under an agreement between the Atomic Energy Commission and the National Oceanic and Atmospheric Administration.

References

Barry, P. J., 1971: Use of argon-41 to study the dispersion of stack effluents. *Proc. of Symposium on Nuclear Techniques in Environmental Pollution*, Int. Atomic Energy Agency, pp. 241-253.

Cramer, H., 1946: *Mathematical methods of statistics.* Princeton Univ. Press, Princeton, N. J.

Gartrell, F. E., 1966: Control of air pollution from large thermal power stations. *Revue Mensuelle 1966 de la Soc. Belge des Ingenieurs et des Industriels Bruxelles*, pp. 1-12.

Gifford, F. A., 1959: Statistical properties of a fluctuating plume model. Proceedings of Symposium on Atmospheric Diffusion and Air Pollution, *Advances in Geophysics. 6:* 117-137.

Hald, A., 1952: *Statistical theory with engineering applications.* Wiley, New York, N. Y.

Larsen, R., 1970: Relating air pollutant effects to concentration and control. *J. Air Pollution Control Association. 20:* 214-225

Larsen, R., 1971: *A mathematical model for relating air quality measurements to air quality standards.* Office of Air Programs, USEPA Pub. No. AP-89.

Prinz, B. and Stratmann, H., 1966: The statistics of propagation conditions in the light of continuous concentration measurements of gaseous air pollutants. *Staub. 26*: 4-12 (English translation edition).

Scriven, R. A., 1965: Some properties of ground level pollution patterns based upon a fluctuating plume model. *CEGB Lab. Note No. RD/L/N 60/65.*

Scriven, R. A., 1971: Use of argon-41 to study the dispersion of stack effluents. *Proc. of Symposium on Nuclear Techniques in Environmental Pollution,* Int. Atomic Energy Agency, pp. 254-255.

Wipperman, F., 1966: *On the distribution of concentration fluctuations of a harmful gas propogating in the atmosphere.* (Translation of unpublished MSS.) 17 p.

DISCUSSION

Arnold Court: Your closing comment that the lognormal does not fit very well at low concentrations is obvious in the derivation. That derivation is not valid when x can approach 0, because you would be dividing by 0 in your derivation. The lognormal will fit only when the fluctuations are small compared to the concentration value.

Gifford: I have no comment; that seems reasonable.

F. B. Smith: Referring back to one of your very first equations. I don't think this equation which relates the concentration at time (t_i) to the concentration at time (t_{i-1}) can be unique. In fact the relationship normally used, for example with wind velocities, is that the velocity, or the concentration in this case, is related to the velocity or concentration of the previous time through a correlation coefficient. In other words, looking at this first equation, x_1 would equal the value x_0 times some sort of decay function, which might be exponential but related to the correlation between the concentrations at those two times, plus some random element, which would have zero mean and some specified standard deviation. I think this would probably give quite a different distribution, a different answer. I'm not quite sure whether it would give a lognormal.

Gifford: Yes, I would be interested to know the answer to this. I think that certainly the important problem is to try to say something more about why, other than just that it's an irregular function. I would certainly think it would be a good idea to use different kinds of generating functions and see what the resulting distributions are.

M. M. Benarie: I really am not here to disprove or attack lognormal distributions, which I use in the next paper to a great extent. All this discussion about the genesis of describing functions reminds me very much of the discussion in aerosol physics which began about 25 years ago, but was luckily ended about 10 years ago, about the exponential, Risen-Rammler, lognormal and other descriptive functions for the distribution of aerosols. For me, any function which is mathematically easy to handle and is a good approximation is good. So why not take the lognormal?

James Arveson: I have two questions - one is for information, the other might open a Pandora's Box. The first one is for your equation 4. Why is there circular symmetry? Why not allow the possibility of some kind of correlation in the model? The second question is why is there a need to have a parametric formulation for these distributions? Perhaps we should be content to deal with just quantiles of generally non-parametric distributions. Why is there the need to fit a lognormal, a Weibull, etc.?

Gifford: This equation 4 doesn't have cross-correlation terms in it simply because, in the usual form, the distributions with respect to y of the concentration is assumed not to involve a cross-product term in σ_y and σ_z. In conditions of strong stability this is undoubtedly a poor assumption and I don't think there is any particularly good reason for not setting up the model on the basis that you suggest; I just haven't done it here. As to the second question, I don't see offhand—it goes back to what the introductory speakers were saying—you want to be able to rationalize what you observe in terms of physical variables. Now the exact details of the model for doing this could be debated, but it seems to me that whether you use...I am trying to see rather desperately, not being a statistician, how what you would call a non-parametric approach would apply here, but it seems to me that it is something like, "Look Ma, no hands." What you want is a way of relating the atmosphere, which is the physical medium, to the observed concentration values, because you need to specify the transfer function in the atmosphere, and I certainly don't have any strong feeling about how that should be done. Here I intended to show only a couple of possibilities that occurred to me.

Ron Snee: I agree that if you had to pick a single distribution function, the lognormal would be the one to pick. I would point out, however, that the Pearson system of distribution functions includes a variety of distribution curves and has been used by statisticians for many years in the characterization of empirical distributions. You did not mention the Pearson system. I wonder if it hasn't been used or if perhaps there is some reason why it shouldn't be used?

Gifford: I don't know of any reason why any rational approximation to what is observed couldn't be used and I certainly didn't intend to imply that it shouldn't be. They're all equivalent. If you, for instance, look at a table of the parameters of distributions, you will find that all of the skew frequency distributions are related. The main difference among the different families of distributions has to do with whether they are discrete or continuous variables. For the rest of it they are mostly more or less all inter-related.

Snee: I would encourage the group to investigate the Pearson system. I believe the Pearson system will help get us out of the problem of deciding whether the lognormal distribution is appropriate in a given instance.

Gifford: The problem isn't the form of the distribution, in my opinion. It's how you go about specifying the physical mechanisms involved, and that's the reason that I don't really care too much about this little explanation here without some rational way of characterizing the y's, which is what Dr. Smith was getting at. The problem is to be able to express the parameters in some suitable distribution in terms of atmospheric physical variables.

4. METEOROLOGICAL PARAMETERS RELEVANT IN A STATISTICAL ANALYSIS OF AIR QUALITY DATA

F. WIPPERMANN

Department of Meteorology
Technical University, Darmstadt, Germany

Introduction

Today many measurements (continuous or at discrete times) of air quality are carried out at many places in industrial countries. Statistical evaluations of these measurements are as different (and therefore incomparable) as the places are. Mostly concentration of a gaseous component or of particulate matter is considered as depending on surface wind (speed and direction), on surface temperature, and sometimes on humidity. However these three or four parameters are not able to describe completely the turbulent state in the atmospheric boundary layer and, therefore, not the diffusion condition in it.

This paper intends to show which meteorological parameters can be considered as relevant in a statistical analysis of air quality data. However the conclusions drawn are valid only if the atmospheric boundary layer satisfies the conditions of a planetary boundary layer (PBL), i.e., stationarity and horizontal homogeneity. In general, since the actual boundary layer differs from the PBL, the conclusions are therefore only approximately correct.

This basic concept has been developed together with Dr. Yordanov (Sofia, Bulgaria) and is the subject of a recent joint paper (Wippermann and Yordanov (1972)).

Planetary Boundary Layer (PBL) and Rossby Number Similarity

The PBL is defined as a steady state horizontally homogeneous boundary layer. In a PBL, there exists, for $z \gg z_0$ (z_0 = roughness length), a so-called Rossby number similarity, which means that the vertical profiles of certain variables (non-dimensionalized correctly by internal parameters) are independent of the given external parameters. They depend only on an internal parameter μ for thermal stratification, and on two internal parameters λ_x and λ_y

for baroclinicity in the PBL. Of course, height above ground as an independent variable has also to be made nondimensional by a scale height H of the PBL.

$$\dot{Z} = z/H \qquad H = \kappa u_*/f \tag{1}$$

where κ is the von Karman constant; $u_* = (\tau_0/\bar{\rho})^{1/2}$ the friction velocity; τ the Reynolds stress; ρ the density; and f the Coriolis parameter. Variables for which Rossby number similarity is valid are for instance

$$P = \kappa(\hat{u} - \hat{u}_g)/u_* \qquad Q = \kappa(\hat{v} - \hat{v}_g)/u_* \qquad T = \tau/(\bar{\rho} u_*^2) \tag{2}$$

$$E = \kappa^2 \epsilon/(u_*^2 f) \qquad K_m = k_m/(H^2 f) \qquad E_x = \overline{\rho (u'')^2}/(\bar{\rho} u_*^2)$$

$$E_y = \overline{\rho (v'')^2}/(\rho u_*^2) \qquad E_z = \overline{\rho (w'')^2}/(\bar{\rho} u_*^2) \qquad E_x \zeta_x(n)$$

$$E_y \zeta_y(n) \qquad E_z \zeta_z(n) \qquad \Gamma = (\hat{\vartheta}_T - \hat{\vartheta})/\vartheta_* \qquad S = (\hat{s}_T - \hat{s})/s_*$$

P and Q are non-dimensional velocity defects; $\hat{u}$ and $\hat{v}$ are velocity components in a coordinate system, the x-axis of which coincides with the direction of the surface stress τ_0 (an internal parameter); and E is the rate of dissipation of turbulent energy; u'', v'' and w'' are the fluctuations, and E_x, E_y and E_z are the three parts of turbulent kinetic energy; ζ_x, ζ_y and ζ_z are spectral density functions with a frequency n. Γ is the non-dimensional difference of temperature to the temperature at the top of the PBL (index T), and S is the non-dimensional difference of water vapor to the water vapor content at the top of the PBL. $\vartheta_* = -q_0/(\kappa \bar{\rho} c_p u_*)$ is a characteristic temperature fluctuation, with c_p the specific heat and q_0 the turbulent heat flux at the ground. $s_* = i_0/(\kappa \bar{\rho} u_*)$ is a characteristic moisture fluctuation, with i_0 the turbulent moisture flux at the ground, i.e., the rate of evaporation.

All the variables listed in Equation 2 form universal vertical profiles, depending only on the three internal parameters, μ, λ_x, λ_y. For instance

$$T = T(Z, \mu, \lambda_x, \lambda_y) \quad \text{for } Z \gg Z_0 \tag{3}$$

where

$$\mu = H/L_* \tag{4}$$

is the internal parameter for thermal stratification with $L_* = -c_p \bar{\rho} u_*^3/(\kappa \beta q_0)$, the Monin-Obukhov stability length; $\beta = g/\vartheta$ with ϑ a reference temperature.

$$\lambda_x = \frac{\kappa^2}{f}\frac{d\hat{u}_g}{dz} \qquad \lambda_y = \frac{\kappa^2}{f}\frac{d\hat{v}_g}{dz} \tag{5}$$

are the two internal parameters for baroclinicity of the PBL. They are internal ones because they contain the components of $d\hat{\mathbf{v}}_g/dz$ in a coordinate system oriented with the x-axis in the direction of the internal parameter π_0.

For all the variables listed in Equation 2 the same as in Equation 3 is valid. This means that the vertical profiles depend only on μ, λ_x, λ_y. This means furthermore that the state of turbulence in the PBL, and, therefore also the turbulent diffusion, is completely described by these three parameters.

The Vertical Profile of Concentration Caused by a Horizontal Surface Source

The condition of horizontal homogeneity of the PBL is satisfied as long as the source of a gas or of particulate matter is a horizontal and infinite surface source. This means that the concentration $\hat{r}_s$ [g cm^{-3}], made dimensionless by the characteristic fluctuation of concentration $r_* = i_0/(\kappa\bar{\rho}u_*)$, with i_0 [g cm^{-2} sec^{-1}] the source strength, must have a universal vertical profile

$$R_s = \frac{\hat{r}_s}{r_*}(Z, \mu, \lambda_x, \lambda_y) \quad \text{for } Z >> Z_0 \tag{6}$$

Actually the Rossby number similarity is valid for the non-dimensional difference $(R_s)_T - R_s$. Here $(R_s)_T$, the concentration at the top of the PBL, is made zero (vanishing background concentration). Examples of natural sources of this kind are an evaporating sea surface or a very large sand desert with sustained strong winds. Water vapor or sand is added to the air. The vertical profile of concentration of these admixtures depends only on the three parameters μ, λ_x, λ_y.

The Concentration in the Case of a Continuous Point Source

If one considers a continuous point source, the condition of horizontal homogeneity is violated and Rossby similarity can no longer be used to conclude on which meteorological parameters the concentration pattern depends. (In the case of an instantaneous point source, the condition of stationarity is also violated).

One can try to make a statement concerning the relevant meteorological parameters by assuming that the diffusion process is described by the steady-state Fickian diffusion equation

$$\hat{u}\frac{\partial \hat{r}_p}{\partial x} + \hat{v}\frac{\partial \hat{r}_p}{\partial y} = \frac{\partial}{\partial x}\left[k_x \frac{\partial \hat{r}_p}{\partial x}\right] \qquad (7)$$

$$+ \frac{\partial}{\partial y}\left[k_y \frac{\partial \hat{r}_p}{\partial y}\right] + \frac{\partial}{\partial z}\left[k_z \frac{\partial \hat{r}_p}{\partial z}\right]$$

and by replacing the velocities of û and v̂ and the turbulent diffusion coefficients k_x, k_y, k_z by variables, for which universal vertical profiles exist. The horizontal coordinates should be made dimensionless by the internal scale height H given in Equation 1b, X = x/H and Y = y/H, and the diffusion coefficients by H^2f.

Since

$$\hat{u}_g(z) = \hat{u}_{go} + \frac{d\hat{u}_g}{dz} z = \hat{u}_{go} + \frac{\mu_*}{\kappa}\lambda_x Z$$

$$v_g(z) = v_{go} + \frac{dv_g}{dz} z = v_{go} + \frac{\mu_*}{\kappa}\lambda_y Z$$

and

$$\hat{u}_{go} = |\hat{w}_{go}| \cos(\alpha_0)$$

$$-\hat{v}_{go} = |\hat{w}_{go}| \sin(|\alpha_0|)$$

there results

$$\hat{u}/(Hf) = \left[P + \kappa \cos(\alpha_0)/C_g\right]/\kappa^2 \qquad (8)$$

$$\hat{v}/(Hf) = \left[Q - \kappa \sin(|\alpha_0|)/C_g\right]/\kappa^2$$

a_0 is the cross-isobar angle and $C_g = u_*/|\hat{w}_{go}|$ the geostrophic drag coefficient. Both α_0 and C_g can be eliminated in Equation 8 by making use of the resistance law for the PBL

$$\kappa \cos(\alpha_0)/C_g = -M_m(\mu, \lambda_x, \lambda_y) + \ln(Ro_0 C_g) \qquad (9)$$

$$\kappa \sin(|\alpha_0|)/C_g = N(\mu, \lambda_x, \lambda_y)$$

where $Ro_0 = |\hat{w}_{go}|/fz_0$ is the surface Rossby number, a non-dimensional combination of given external parameters. The functions N and M_m appearing in this law are universal functions, i.e. they are independent of external parameters and depend only on μ, λ_x, λ_y. The resistance law was first derived by Kazanskii and Monin (1961) for the barotropic and neutral case. It was extended to the diabatic case by Blackadar (1967) and by Monin and Zilitinkevich (1967). Recently it was extended to baroclinic cases by Yordanov and Wippermann (1972). By using the resistance law Equation 9a and 9b and the definition

$$Z_0 = (\kappa \, Ro_0 \, C_g)^{-1} \qquad (10)$$

one obtains for the diffusion Equation 7, the following form

$$(P - M_m + \lambda_x Z)\frac{\partial R_p}{\partial X} + (Q - N + \lambda_y Z)\frac{\partial R_p}{\partial Y} - \ln(\kappa Z_0)\frac{\partial R_p}{\partial X} = \kappa^2 \left\{ \frac{\partial}{\partial X}\left[K_x \frac{\partial R_p}{\partial X}\right] + \frac{\partial}{\partial Y}\left[K_y \frac{\partial R_p}{\partial Y}\right] + \frac{\partial}{\partial Z}\left[K_z \frac{\partial R_p}{\partial Z}\right]\right\} \qquad (11)$$

The boundary conditions are

$$\begin{array}{llll} X \rightarrow \infty, & Y \rightarrow \pm\infty, & Z \rightarrow \infty: & R_p = 0 \\ X = 0, & Y = 0, & Z = Z_b: & R_p = 1 \\ Z = Z_0: & & & \partial R_p / \partial Z = 0 \end{array} \qquad (12)$$

$R_p = \hat{r}_p/(b \, f^5 \, g^{-3})$ is the non-dimensional concentration caused by a continuous point source; b [g sec^{-1}] is the strength of the source; and $Z_b = z_b/H$ is the non-dimensional effective height of the source. If one assumes that the vertical profiles of the non-dimensional diffusion coefficients for matter K_x, K_y, K_z are universal ones like the vertical profile of the diffusion coefficient K_m for momentum, all coefficients in Equation 11—except Z_0—depend only on μ, λ_x and λ_y and, of course, some of them on the independent variable Z. They all are universal functions. However the non-dimensional roughness length Z_0 depends on Ro_0 and C_g as seen in Equation 10, and C_g itself depends on Ro_0, μ, λ_x, λ_y, as seen in the resistance law, Equations 9a and 9b. Therefore since Z_0 depends on Ro_0, μ, λ_x, λ_y the non-dimensional concentration R_p must also depend on Ro_0.

$$R_p = R_p(X, Y, Z, Z_b, Ro_0, \mu, \lambda_x, \lambda_y) \qquad (13)$$

Universality is now lost; dependence from Ro_0 (i.e., from external parameters) enters because of violation of the condition of horizontal homegeneity by having a continuous point source. A comparison of R_p in Equation 13 with R_s in Equation 6 shows the difference.

The Concentration (at a Fixed Point) Caused by Multiple Sources

A statement can be made only if one assumes that the sources do not change their coordinates and their strengths during the period of measurement. Furthermore one has to assume that the period of measurement is long enough to cover all possible cases (wind direction, Rossby number, thermal stratification and, possibly, the baroclinicity) in almost equal parts. This later assumption has to be made in almost all statistical analyses of measured concentrations. However, the first assumption will be only incompletely fulfilled and therefore causes errors.

For each direction (of the geostrophic surface wind) the mean concentration $\bar{r}_d$ at the point under consideration should be evaluated and a non-dimensional concentration

$$R_d = \hat{r} / \bar{r}_d \tag{14}$$

should be formed therefrom. Fluctuations of R_d should be independent of the source positions $(x_b)_i$, $(y_b)_i$, $(z_b)_i$ and of the source strengths b_i. They should depend on the remaining parameters in Equation 13.

$$R_d = R_d\,(Ro_0, \mu, \lambda_x, \lambda_y) \tag{15}$$

The surface Rossby number $Ro_0 = |\hat{\mathbf{v}}_{go}|/(fz_0)$ can be determined directly from the given geostrophic surface wind, when z_0 in the denominator is known. This can be obtained by conventional measurements of the wind profile near the ground, but should be representative of the whole area in which diffusion takes place. It may vary with the wind direction (and is therefore a "meteorological" parameter) and, possibly, with the vegetation period.

The internal parameter μ for thermal stratification has to be found by converting the external parameter σ for thermal stratification

$$\sigma = D / \Lambda \tag{16}$$

into the internal one, where D is the external scale height of the PBL and Λ is the external stability length of the PBL

$$D = \kappa^5 \left| \hat{\mathbf{v}}_{go} \right| / f \tag{17}$$

$$\Lambda = \kappa^3 \left| \hat{\mathbf{v}}_{go} \right|^2 / (\beta \delta \hat{\vartheta})$$

with $\delta\hat{\vartheta} = \hat{\vartheta}_T - \hat{\vartheta}_0$. When the temperature difference $\delta\hat{\vartheta}$ from top to bottom and $\hat{\mathbf{v}}_{go}$ the geostropic wind at the ground are given, parameter σ then can be formed. This parameter must be converted to parameter μ. Zilitinkevich (1970) gives diagrams for the conversion of given external parameter σ into the wanted internal parameter μ. For this conversion Ro_0 is again needed.

Of course difficulties will arise in forming external parameter σ from the given parameters ϑ_T, $\hat{\vartheta}_0$, $\hat{\mathbf{V}}_{go}$, because the first two are difficult to find. A PBL can only have a monotonically decreasing or increasing temperature profile $\hat{\vartheta}(z)$. It does not know temperature inversions or layers with unstable stratification or similar things very frequently observed in the real atmospheric boundary layer. Therefore an observed temperature profile has to be smoothed in order to obtain the corresponding profile belonging to the PBL. The difference $\vartheta_T - \vartheta_0$ should be taken from such a profile.

The baroclinicity seems to be less important. It has to be considered only for muchly elevated sources, e.g., very tall stacks. An example has been given by Wippermann and Yordanov (1972) showing a case with a pronounced minimum of eddy diffusivity in 320 m caused by baroclinicity. When baroclinicity must be considered, one has first to form the two external parameters (in a coordinate system x^*, y^* oriented with the x-axis in the direction of the geostropic wind at the surface)

$$\eta_{x*} = \frac{\kappa^2}{f z_T} \left\{ (\hat{\mathbf{V}}_{go})^0 \cdot \hat{\mathbf{V}}_{gT} - \left| \hat{\mathbf{V}}_{go} \right| \right\} \tag{18}$$

$$\eta_{y*} = \frac{\kappa^2}{f z_T} \mathbb{k} \cdot \left[(\hat{\mathbf{V}}_{go})^0 \times \hat{\mathbf{V}}_{gT} \right]$$

where z_T is the height of the top of the PBL. For a conversion into the internal parameters λ_x, λ_y of baroclinicity, the cross-isobar angle α_0 is needed:

$$\lambda_x = \eta_{x*} \cos(\alpha_0) + \eta_{y*} \sin(|\alpha_0|) \tag{19}$$

$$\lambda_y = -\eta_{x*} \sin(|\alpha_0|) + \eta_{y*} \cos(\alpha_0)$$

The angle α_0 can be obtained from the resistance law Equations 9a and 9b.

Concluding Remarks

The baroclinicity of the boundary layer flow has an effect on the concentration pattern only when these are caused by much elevated sources; the effect can be neglected in most cases. The remaining meteorological parameters are the internal parameter μ for the thermal stratification of the boundary layer and the surface Rossby-number $R0_0$. Both these parameters determine the concentration uniquely (as long as the assumptions of a PBL are satisfied).

It seems that the parameter μ is just the one which has been sought as a measure of diffusion characteristics which depend mainly on thermal stratification. The empirical "dispersion categories" can possibly be replaced by this parameter, if we succeed in determining the PBL which is equivalent to the (measured) actual boundary layer.

References

Blackadar, A. K., 1967: External parameters of the wind flow in the barotropic boundary layer. *GARP Study Conference Report,* Stockholm, Sweden, Appendix IV, 11 p.

Kazanskii, A. B., and Monin, A. S., 1961: On the dynamic interaction between the atmosphere and the earth surface. *Bull. (Izv.) Acad. Sci. USSR, Geophys. Ser. Nr. 5,* 786-788, English translation. 514-515.

Monin, A. S., and Zilitinkevich, S. S., 1967: Planetary boundary layer and large scale atmospheric dynamics. *GARP Study Conference Report.* Stockholm, Sweden, Appendix V, 37 p.

Wippermann, F., and Yordanov, D., 1972: A perspective of a routine prediction of concentration patterns. *Atmospheric Environment. 6:* 877-888.

Yordanov, D., and Wippermann, F., 1972: The parameterization of the turbulent fluxes of momentum, heat and moisture at the ground in a baroclinic planetary boundary layer. *Beitr. Phys. Atm. 45:* 58-65.

Zilitinkevich, S. S., 1970: The dynamics of the atmospheric boundary layer. *Gidrometeor.,* Leningrad, (in Russian) pp. 291.

DISCUSSION

Smith: Could I ask you, Dr. Wipperman, if you consider that the depth of the boundary layer could be adequately given by the Rossby similarity theory. My experience with this is that one can use the theory very adequately to give you estimates of the surface stress and the turning of the wind at the surface, but in unstable conditions it doesn't normally give very good estimates of the depth of the boundary layer which you've used in your scaling. Usually the depth of the boundary layer depends much more on the historical development of the boundary layer due to the input of heat over the daytime period.

Wipperman: This is a question of how one defines the depth of the boundary layer. So if you have, for instance, an inversion on the top and you consider the height of this inversion as depth of the boundary layer, this could not be considered in a planetary boundary layer in which an inversion is not possible. I'm just choosing this "H" as a scale height for the boundary layer, which does not mean somewhere is the top of this boundary layer.

5. THE USE OF THE RELATIONSHIP BETWEEN WIND VELOCITY AND AMBIENT POLLUTANT CONCENTRATION DISTRIBUTIONS FOR THE ESTIMATION OF AVERAGE CONCENTRATIONS FROM GROSS METEOROLOGICAL DATA

M. M. BENARIE

Institut National de Recherche Chimique Appliquee
Vert-le-Petit, France

Introduction

A synonym for "computation of pollutant concentrations from meteorological data" is atmospheric modeling. In this matter, one has on the one hand mechanistic (or explaining) models, which seek the breakdown as well as the comprehension of the elementarv physical processes of dispersion. On the other hand, one has formal (or phenomenological) models, which look for the necessary and sufficient coefficients for the computation of some mean or probability of given concentrations. After explaining the reasons why I have not chosen a mechanistic model for concentration frequency computation, I will deal further with one specific model.

This distinction is very near to the one defined by Stern (1970) at this place just two years ago, in his Symposium Summary on Multiple-Source Urban Diffusion Models: mechanistical models are source-oriented and the phenomenological ones are receptor-oriented. It should nevertheless be stressed that phenomenological and statistical models do not necessarily mean the same thing. Mechanistical (source-oriented) models are constrained by considerations of material balance, as opposed to statistical (empirical) ones, which are not (Calder (1970)). The receptor-oriented phenomenological model proposed herewith does not implicitly avoid the use of the law of the conservation of matter (even if in the present paper we fail yet to attain its ultimate consequences) and it has the pretension to give more insight into physical processes than just a correlation between measured pollutant concentrations and simultaneously observed meteorological parameters.

Concentration frequency distributions are generally obtained by calculating the concentration values for all possible combinations of meteorological parameters: wind direction, wind speed and stability category. Afterwards, we take the sum of the joint frequencies of all combinations of classes that rise to a given concentration. The first objection to this way of proceeding is economical.

Since the chosen dispersion equation has to be evaluated numerically at least a few hundred times for each receptor location of any interest, such extensive computation can quickly involve prohibitive time, even with a high-speed computer.

Secondly, the relevant meteorological data or statistics have to be extensively known for the given location. While these data may be available in some cases, in others not less important, the next meteorological station—perhaps a hundred miles away—may not be at all representative.

Thirdly, the result of individual computations of the dispersion equation has the character of a differential. In the case of important point sources, the result justifies careful consideration of the constants, which are often based on extensive surveys of emission and meteorological parameters. As maximum concentrations are mostly sought, the results are acceptable even when off the target by a factor of two or even more. But if concentration frequency distributions, or one of their derivatives as a mean, are sought, the input errors in the dispersion equations (including lack of basic meteorological information) can easily be amplified by the effect of the summation. As experience shows, the final result is, as often as not, off by ±50%. This error is unacceptable since any air pollution engineer, worthy of this name, should be able to estimate in ten minutes on a slide rule from very few data (such as population density, space heating habits, industrial context and some general knowledge about climatology), a mean with the specified accuracy.

The fourth objection that can be made is that the basic concept of plume computations is a short-time process, say typically of 30 minutes. The passage to long-time averages is especially awkward at the level of transition from gaussian plume of *constant* direction to the normally meandering plume. Next follows another transition to changing wind directions, as described by a specific wind rose. At least two different physical processes are involved, one of which is definitely not gaussian.

The fifth and last, but not least, objection is that a model should require a minimum of basic assumptions and discard everything not absolutely needed. This should apply to stability classes, as far as averages are concerned. Not that a frequency table of the stability classes for towns and most populated areas should be so difficult to obtain. However such tables are usually unavailable when and where they are most needed, e.g., for the location of a new plant (see also the second argument above).

And thus arises the question as to whether frequency tables of stability classes are necessary. Looking at Table I, we are inclined to answer that, at least for the limited purpose of computation of averages, they are not absolutely indispensable.

It would be highly interesting to supplement the data of Table I (the only ones we have been able to locate), with non-European statistics and with figures concerning other than temperate climates. From what is available for the time being, we can see that the near-neutral classes represent 76% ±5% of all situations*. We effectuated a few trial computations which show that the mean is but slightly—and the median not at all—sensitive to the frequency changes in the stable and unstable classes. If this is the case, and the overall class frequency distribution is an approximately constant property of the temperate climates, then why bother to split up first into categories and afterwards totalize them during expensive computer hours?

The purpose of the present paper is to provide a simplified method of estimation of pollutant concentrations, in cases where detailed meteorological data are not at hand. As far as possible, the method is an empirical one and has no pretension to give insight into the physical processes of atmospheric dispersion. The purpose is to provide an easy and rapid means for atmospheric modeling around point sources.

The Experimental Data

The experimental basis for the present work is due to Prof. P. Bourbon and his staff who obligingly permitted us to make the following statistical analysis of the data of their survey around the natural gas sweetening plant at Lacq (southwest France). Gratitude is expressed at this point for this important contribution.

For several years, 24-hour mean concentrations of SO_2, NO_2, H_2S and other pollutants were measured at 37 points. Figure 1 shows their repartition around the plant. For the time being, we will use only the data for the two years, 1968-1969, which are complete and homogeneous insofar as SO_2 and NO_2 concentrations are concerned. The former was analysed by the very specific nitroprussate-pyridine method (Bourbon et al. (1971)); the latter by the Jacobs-Hochheiser method.

The concentrations were represented in the form of cumulative frequency diagrams, one for each survey point and pollutant. Of the 74 diagrams, 54 are very nearly straight lines on the lognormal coordinates which were used. In Figure 1 the survey points with two straight-line distributions are marked⌖; those with one by-○-; Figures 2, 3, 4 and 5 give samples of these distributions.

Discussion

It has been found that the cumulative frequency distributions of suspended particulates at CAMP (urban) sites have a tendency toward lognormality (U.S.D.H.E.W. (1958)). Earlier applications, referring also to other pollutants but always to receptors located in or at area sources, are to be found in Zimmer et al. (1959) and Gould (1961). As stated by Gifford (1969), the lognormal concentration distribution can be mathematically derived by the particularization of the general explanation of skew distributions.

---

*) With the restrictive condition that classification criteria should be the same. This meant that we had to leave the Szepesi (1964) data out of Table I because they are based on a perhaps better but nevertheless different criterion, i.e., the measurement of the temperature gradient.

It was shown by Benarie (1969) (1971): (a) that the lognormal distribution is strictly valid for concentration frequency distributions in any given direction around a *point* source. This is a consequence of the facts that wind velocity distributions in any given direction may be approximated by a lognormal and of some very general mathematical properties of this function.* (b) that in the special case of the point source *without thermal plume rise,* the geometrical standard deviations of the wind velocity distribution and that of the concentration distribution are numerically equal. From this equality it follows as a corollary: the concentration frequency distributions for receptors, situated at various distances along the same radius and from the point source, should have the same geometrical standard deviation. The general case of the *source with plume rise,* will be discussed further when speaking about SO_2 (c) that the observed lognormal distributions for *area* sources follow directly by summation of the effect of a large number of likely distributed point sources.

At this point, we should distinguish between NO_2 emitted at nearly ambient temperature and the SO_2 contained in plumes of higher temperature. The former contributes evidence to points (a) and (b) above. As these general affirmations were obtained from relatively few data, this further evidence is useful. The discussion of the SO_2 results below will add a new contribution.

Figures 2, 3, 4 and 5 illustrate the above affirmations (a) and (b). Geometric standard deviations for (unperturbed) NO_2 receptors in the same radial orientation are identical and nearly equal to the geometrical standard deviation of corresponding wind velocity just as required by the corollary to the theorem cited in the Appendix.

Figures 2, 3, 4 and 5 are only a fractional sample of the evidence on hand. Although they are quite convenient for interpolation, as will be shown below, space does not permit displaying similar figures for all 37 survey points. Instead, the principal information from them have been summarized in Figure 6, which displays values of $k_{NO_2} = \sigma_{NO_2} / \sigma_w$ (σ_{NO_2} and σ_w are the respective geometrical standard deviations for NO_2 concentrations and wind velocity). Values of k_{NO_2}, deviating from unity, are found in the western half of the pattern, where topographical accidents are more pronounced. In the eastern half, in the first approximation level the behaviour of k_{NO_2} is as theoretically expected.

As for the SO_2 which is definitely associated with a thermal plume rise and is emitted by 60 to 80 m high stacks, the hypothesis of (concentration) prop w^{-1} cannot be assumed and a $k_{SO_2} = \frac{\sigma_{SO_2}}{\sigma_w}$ value different from unity should be expected. The numerical value of k_{SO_2} may readily be computed by applying a dispersion formula to a plume rise expression. As one has a rather wide choice from both sort of equations and naturally an even larger one of combinations, it is easy to find one or more to "prove" that the k_{SO_2} values reported in Figure 7 are correct. For us, as far as they were empirically observed, they are indeed.

---

*See Appendix, for discussion of the frequency distribution of wind velocity and the properties of the lognormal distribution which are of interest here.

It has already been mentioned that cumulative frequency diagrams such as Figures 2, 3, 4 and 5 are convenient for interpolation purposes. At least for level terrain, the sequence of (almost) parallel, straight lognormal representations is related to distance. This is easy to understand, as concentrations diminish with distance, or, what is equivalent, a given concentration occurs with decreasing frequency and with increasing distance.

This relation between concentration at constant frequency and distance, is illustrated by Figure 8. As most survey points show some (topographical) singularity, it is not easy to align enough points, in order to judge the form of the regression and the exactitude of fit of some function. Therefore Figure 8 should be considered as an empirical data collection and will be used in the following as such, as means for interpolation.

We may now investigate whether there is a correlation for a given distance between the frequency of exceeding some concentration and the frequency of wind blowing from the source in the direction of the receptor.

Figure 9 is the wind rose observed between 1961 and 1965 near the plant site. Frequencies corresponding to the opposite wind directions are the abscissae of Figure 10, concerning only receptors at approximately the same distance, between 5 and 7 km, in this case. The ordinates are the frequencies by which some given concentration—here 50µg NO_2/m^3—are exceeded. It might be expected that a correlation should exist between these frequencies. At first approximation, this assumption seems to be verified.

The observed scatter is due, among other causes, presumably to the lateral wind turbulence and its directional change during a 24-hour sampling period. Probably, with shorter sampling times, this scatter would diminish.

Example of Application

Up to this point, we have presented these experimental data somewhat differently from the usual tabular or isoconcentration-map form. How far reaching is this special presentation? Should it be called a model; or, more modestly, a relationship between wind and ambient concentration.

Suppose we ask for a cumulative concentration frequency diagram for the point marked X on Figure 1, a location at which no receptor was operated. From Figure 9 it can be seen that the frequency of wind blowing from the stacks in the direction of the receptor is 5.5%. Entering Figure 10 at this abscissa value, the frequency of 25% is read at the ordinate. This *would* be the frequency of exceeding 50 μg NO_2/m^3, *if* the receptor were 6±1 km distant from the source. Actually it is ≈ 3.5 km from the source. The concentration corresponding to this distance (and direction) is interpolated as shown in Figure 8. Thus 62 μg NO_2/m^3 is found. This pair of values (62 μg NO_2/m^3, 25%) is one point of the cumulative frequency diagram. As its geometric standard deviation should be equal to that of the wind, the concentration distribution at this hypothetical receptor is defined.

For a thermal source, the same procedure should be followed except that an experimental k-value should also be determined and stack height, effluent temperature and velocity also taken into account. In our case this is defined only by Figure 7 and therefore cannot be considered of general validity. Nearer details will be supplied in a subsequent paper.

Outlook

These interpolations and perhaps slight extrapolations have limited uses in a dense survey network, as the one just discussed. However the empirical relations (a) wind direction frequency versus frequency of exceeding an arbitrary concentration (Fig. 9) and (b) concentration versus distance (Fig. 8) are generally established. Then a few points per diagram, perhaps three, will be sufficient to obtain concentration versus frequency diagrams for a multiplicity of geographically scattered points. The only additional information needed, are wind roses and frequency distributions of wind velocities. This seems true for plane, undisturbed topography. The evidence under consideration is just enough to say that topographic relief *does* something to the constants. But for the time being, we are unable to express this effect in a general and quantitative way.

If, with more evidence at hand, the functional form and the general constants of both these relations can be found, we shall have a modeling method which will need very little meteorological input. In this way, computer time can be replaced by equivalent graph reading time, and, what is even more important, results will be of irrefutable empirical character. Its advantage, above purely statistical models, is a greater generality, as cause and effect are more evidently related in the present model. We hope to continue working in this direction.

Acknowledgement

I wish to express my gratitude to Mr. P. Bessemoulin and Mrs. T. Menard for all their help, computations, computer programs, tedious graph drawing, etc. involved in the present work.

Table 1. FREQUENCIES OF STABILITY CLASSES

Frequency %						
Un-stable A,B	Near-neutral C,D,E	Stable F,G	Principle of classification	Year	Reference	Site
4.2	76.8	19.0	Pasquill-Turner	1961-62	Nester (1966)	Frankfort, G.
10.1	69.6	20.3	Pasquill	?	Bryant (1964)	? , Br.
5.0	85.7	6.7	t measured	1958-63	Szepesi (1964)	Budapest, Hung.
14.3	70.6	15.1	Pasquill	1964	Polster-Vogt (1965)	Julich, G.
10.2	82.2	7.6	Pasquill-Turner	1965-70	Hodin	Trappes, F.
7.8	81.4	10.7	Pasquill	1967-71	Bessemoulin (1972)	Rouen, F.

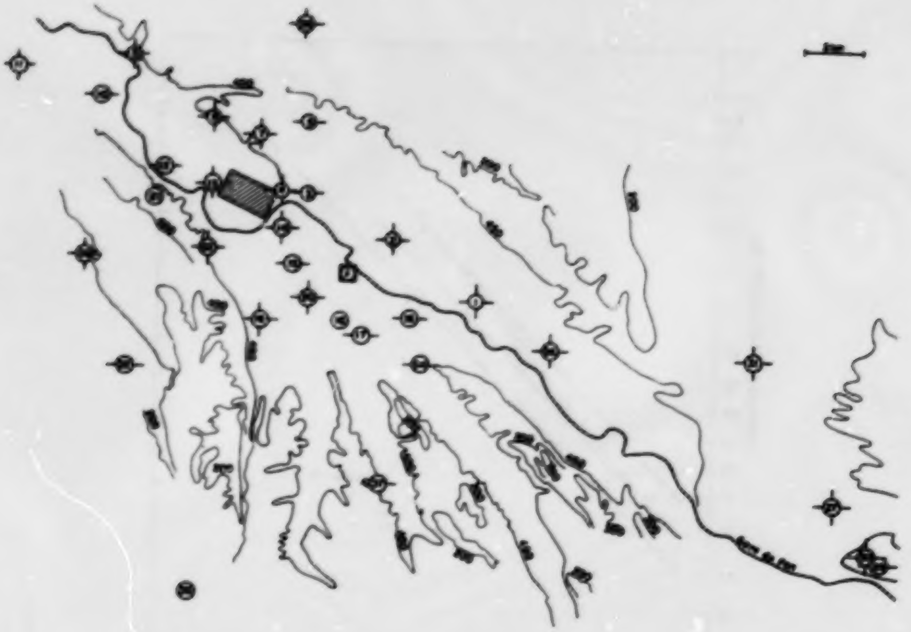


Figure 5-1. Location of sampling stations around gas sweetening plant at Lacq, France.

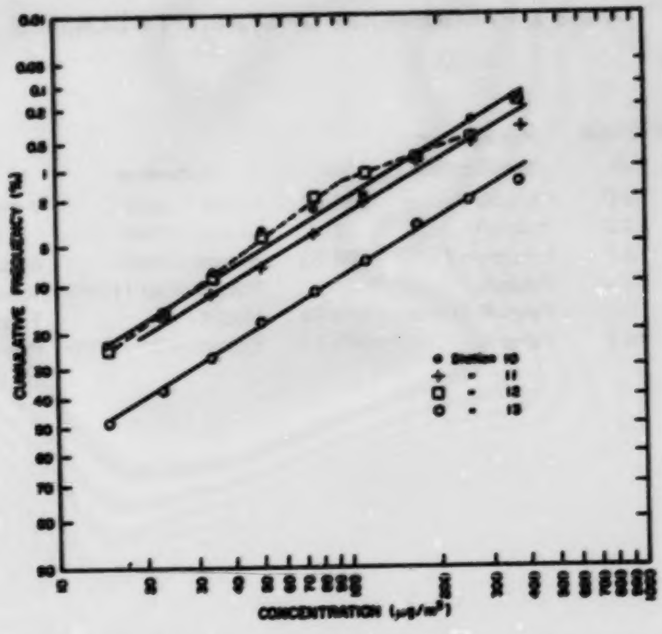


Figure 5-2. Cumulative frequency diagram for SO_2 concentrations at stations located in the 289°-308° sector from the source, at Lacq, France.

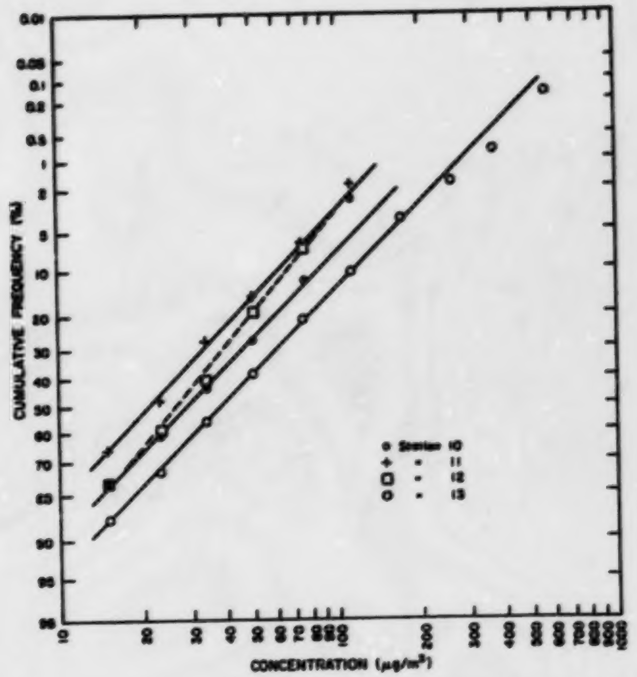


Figure 5-3. Cumulative frequency diagram for NO_2 concentrations at stations located in the 289°-308° sector from the source, at Lacq, France.

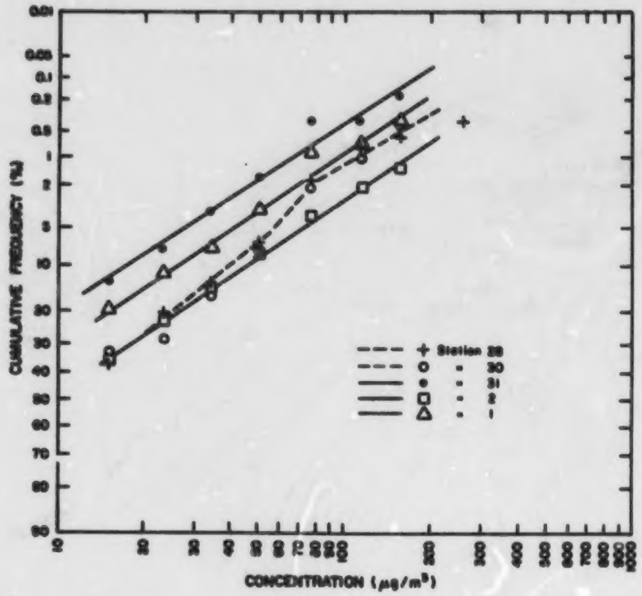


Figure 5-4. Cumulative frequency diagram for SO_2 concentrations at stations located in the 108°-121° sector from the source, at Lacq, France.

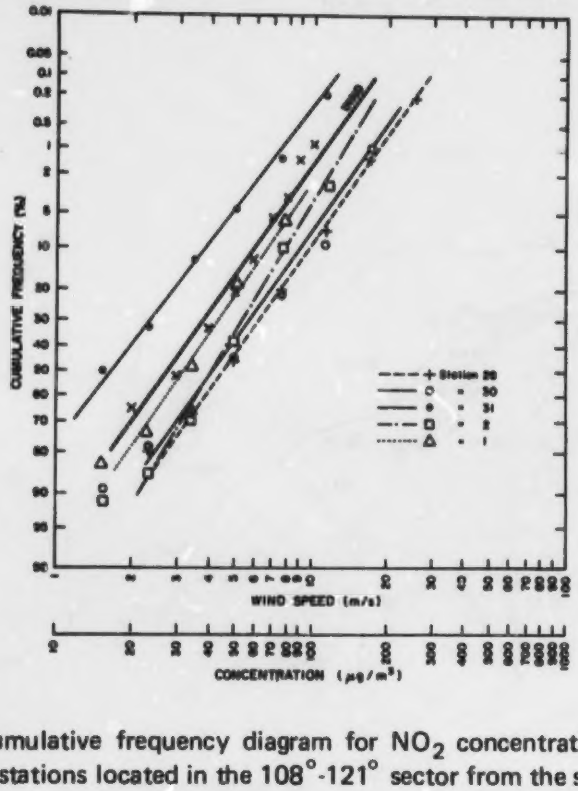


Figure 5-5. Cumulative frequency diagram for NO_2 concentrations and wind speed at stations located in the 108°-121° sector from the source at Lacq, France.

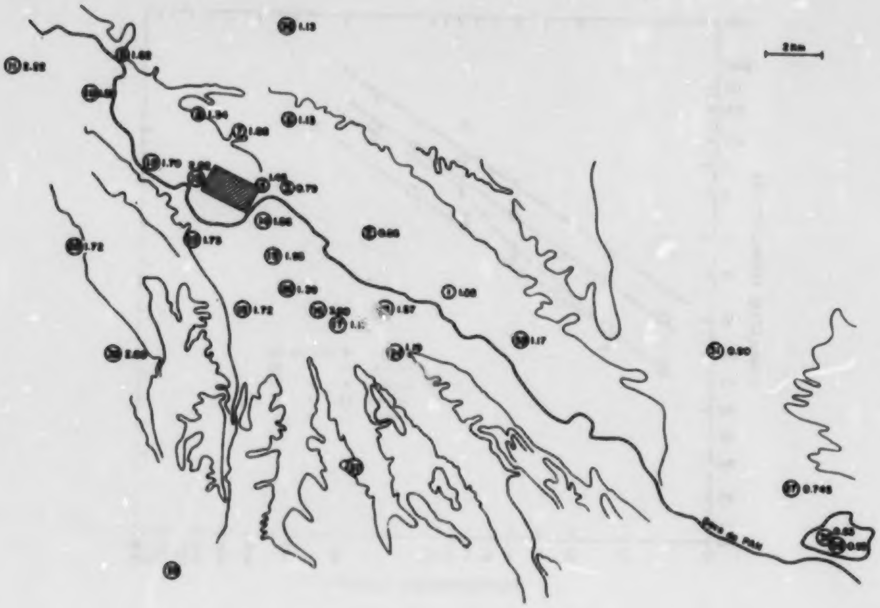


Figure 5-6. Values of $k_{NO_2} = \frac{\sigma NO_2}{\sigma \text{wind}}$ for sampling stations at Lacq, France.

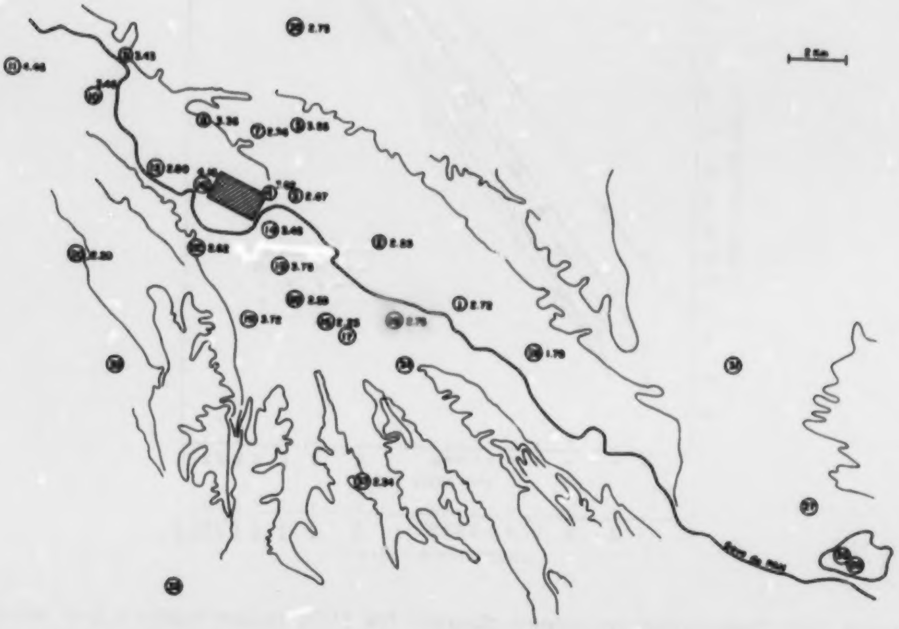


Figure 5-7. Values of $k_{SO_2} = \frac{\sigma SO_2}{\sigma \text{wind}}$ for sampling stations at Lacq, France.

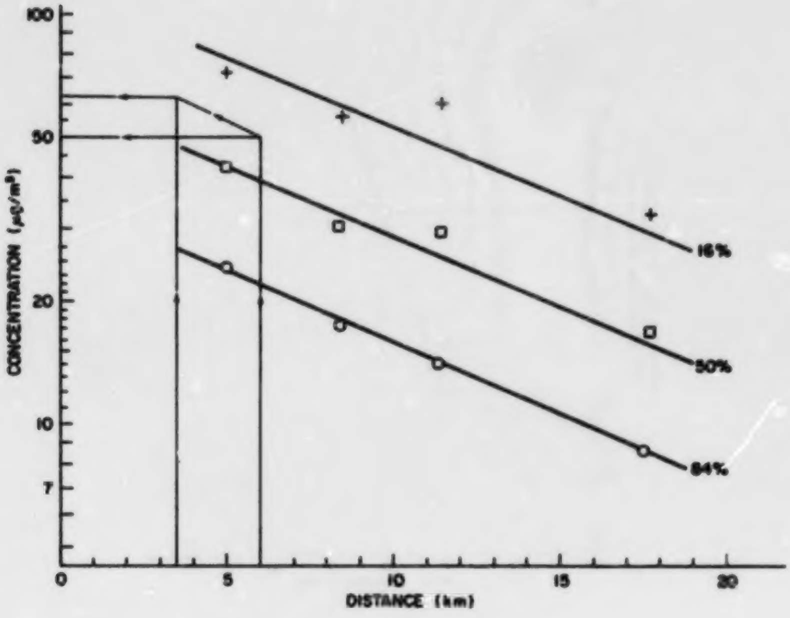


Figure 5-8. The concentration at three frequencies as a function of distance from the source.

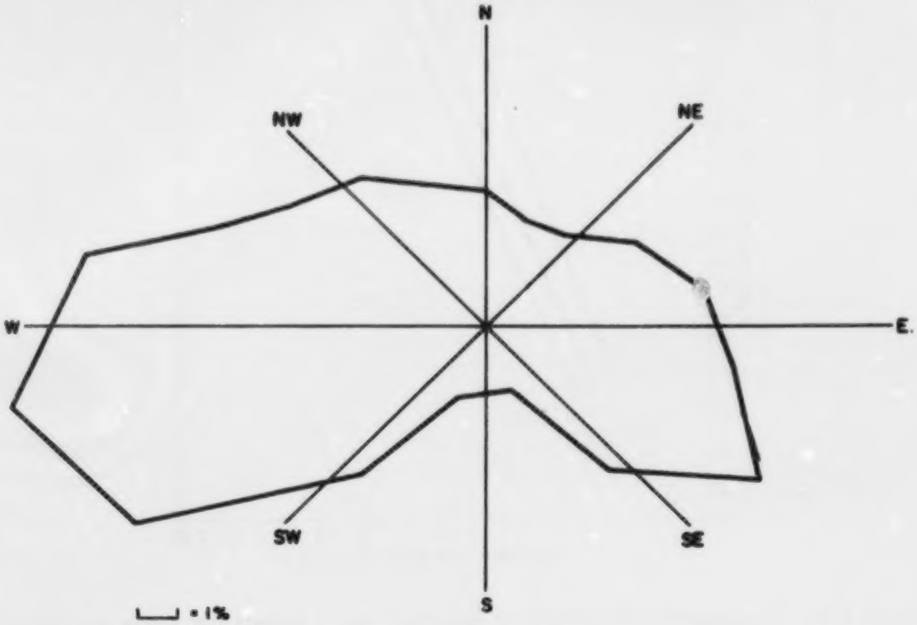


Figure 5-9. Wind rose for 1961 - 1965 at plant site in Lacq, France.

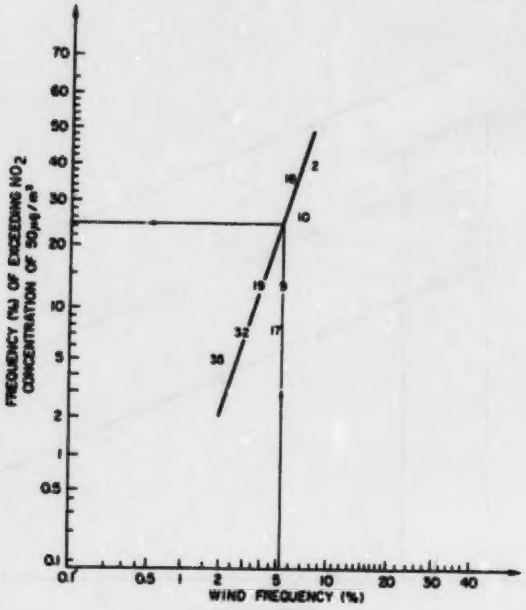


Figure 5-10. Frequency of exceeding 50μg NO_2/m^3 as a function of wind frequency from the source to the receptor for eight sampling stations.

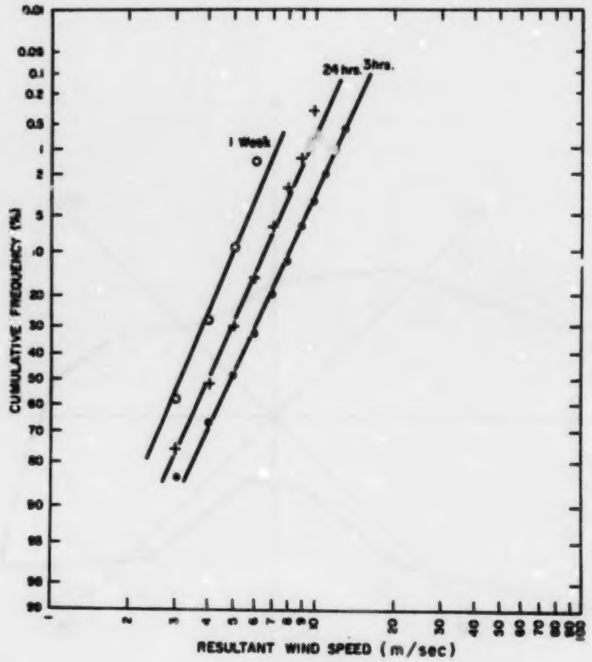


Figure 5-11. Cumulative frequency of resultant wind speeds for three averaging times.

References

Aitchinson J. and Brown, J.A.C., (1969): *The lognormal distribution.* Cambridge University Press, p. 9 ff.

Benarie, M., (1969): Le calcul de la dose et de la nuisance du polluant émis par une source ponctuelle. *Atmospheric Environment. 3*: 467.

Benarie, M., (1971): Sur la validite de la distribution logarithmico-normale des concentrations de polluant. *Proceedings of the 2nd Internat. Clean Air Congress,* 1970, Washington, D. C., Academic Press, New York, N. Y., pp. 68-70.

Bessemoulin, P., (1972): *Contribution a l'etude de la diffusion des polluants gazeux dans l'atmosphere.* Thesis, Paris, France.

Bourbon, P., Malbosc, R., Bel, M. J., Roufiol, F. and Rouzaud, J. F., (1971),: Contribution a la détérmination spécifique dans l'atmosphere du dioxyde de soufre. *Poll. Atm. 52:* 271-275.

Brooks, C.E.P. and Carruthers, N., (1953): *Handbook of statistical methods in meteorology*, Her Majesty's Stationery Office, London, England, chapters 3 and 11.

Bryant, P. M., (1964): Methods of estimation of the dispersion of windborne material and data to assist in their application. *AHSB (RP) R42.*

Calder, K. L., (1970): Some miscellaneous aspects of current urban pollution models. *Proc. Symposium on Multiple-Source Urban Diffusion Models,* Research Triangle Park, N. C., APCO Pub. No. AP-86.

Gifford, F. A. Jr., (1969): The lognormal distribution of air pollution concentrations. Air Resources Atmospheric Turbulence and Diffusion Laboratory, ESSA, Oak Ridge Tennessee (preprint, 3p.).

Gould, G., (1961): The statistical analysis and interpretation of dustfall data. *Proc. 54th Annual Meeting Air Pollution Control Association,* New York, N.Y.

Hodin, M.: Personal Communication.

Nester, K., (1966): Haufigkeitsstatistische Aussagen uber Maximalkonzentrationen von Schornsteinabgasen auf Grund synoptischer Wetterbeobachtungen. *Staub. 26:* 521.

Polster, G. and Vogt, K. J., (1965): Grundsatze and Untersuchungen zur Beurteilung der Ausbreitung radioaktiver Abluft. Protokoll zur Informations- und Arbeitstagung im Kernforschungszentrum Karlsruhe, Germany, p. 15.

Stern, A. C., (1970): Utilization of air pollution models. *Proc. of the Symposium on Multiple-Source Urban Diffusion Models,* Research Triangle Park, N. C., APCO Pub. No. AP-86.

Szepesi, P., (1964): Computations of concentrations around a single source. *Idojaras. 68:* 257.

U.S.D.H.E.W., (1958): Air Pollution Measurements of the National Air Sampling Network - Analyses of Suspended Particulates, 1953-57. *PHS Publication No. 637.* p. 245.

Zimmer, C. E., Tabor, E. C. and Stern, A. C., (1959): Particulate pollutants in the air of the United States. *J. Air Pollution Control Association. 9:* 136.

Appendix

It is fairly well known in meteorology (Brooks, et al. (1953)) that the distribution of wind velocities is skew in a given direction, with high frequencies at low velocities. Several two or more parameter laws present a fair approximation of the experimentally observed distributions.

It has been observed that among two-parameter skew distributions the logarithmic normal function is an experimentally convenient representation of the wind velocity (Benarie (1969)).

At first, it seems singular to use a mathematical function which will not accomodate the zero value of the variable for wind. Measured with the usual cup anemometer, wind velocity values almost everywhere show a high percentage of "calm" periods. Closer scrutiny of sensitive thermoanemometric data seems to suggest that this abundance of calms is purely instrumental. In reality, very low velocities occur with finite frequencies, and a true zero does not physically exist. As our present purpose is not to get into meteorological arguments, we avoid this difficulty by defining wind velocity classes as "less than 1 m/sec." (the starting point of the anemometer) and by including in this class all observations between 0 and 1 m/sec. The fraction of observed "calms" is proportionately attributed to each directional frequency.

A second difficulty arises from the fact that (except for specially conceived survey networks which we do not possess) wind data are from meteorological stations, following international meteorological conventions (i.e., one observation every three hours.); while pollutant concentration data are integrated for shorter or longer periods (24 hours in our case). Figure 11 which presents the cumulative frequency distributions of: (a) 3-hour, (b) 24-hour and (c) 1-week wind vectors from the same station, shows that the error committed by using (a) instead of (b) is slight. Anyway, this is a minor point, as the 24-hour wind vector, which should be physically better justified, can be obtained easily from the original data by a minor computational program.

This rather lengthy argument about the approximation of the observed distribution of wind velocity frequency by a lognormal function was necessary because of the interesting reproductive properties of this two-parameter distribution (Aitchinson and Brown (1969)), which are the immediate consequences of those for the normal distribution.

Theorem: if w is $\Lambda(\bar{w},\sigma)$ i.e. a lognormal function with the geometric mean $\bar{w}$ and the geometric standard deviation σ, and k and c are constants, where $c > o$ (say $c = e^a$), then cw is $\Lambda(a + b\bar{w}, k\sigma)$.

This theorem implies the corollary result: if w is $\Lambda(\bar{w},\sigma)$ then w^{-1} is $\Lambda(-\bar{w},\sigma)$.

DISCUSSION

Donald Rote: I'm not completely sure but it seems to me that your approach depends very heavily upon having a uniform wind field. If you have topographic features that in some way influence the wind field, you will have differences at the same time between wind directions at the source, and at a given receptor. As a consequence, this will greatly distort your capability of generating curves of constant percentile. Could you comment on that please.

Benarie: I fully agree with you. The fact, in Figure 7 I think, of having ratios different from the expected value of 1 in the western part of the pattern, and about 1 in the eastern part, which is level, illustrate your point very well. But survey data are very expensive; I had to do with the data I had and these were the many meteorological data I had. The correct experiment to verify would be to have had wind vanes at at least 8 stations. Then the conclusion would be immediate or almost immediate. I agree fully with your point.

Harold Neustadter: Have you had an opportunity yet to attempt any internal check on the validity of your conclusion. Namely taking three or four of your receptors and seeing if you can generate the results of your dense network within the set you already have?

Benarie: Sure, I did. That was the first check, and it was as reliable as the receptors and measurement results. You know, of course, that manual chemical and analytical methods are good to, say, plus or minus 20 percent. I can't pretend more.

Singer: Work like this is being done by Brookhaven National Lab where they are studying a network outside of New York City. A paper was just presented by Gil Raynor at the Philadelphia meeting where he had concentration vs. distance from New York City out to a hundred kilometers and it is very similar to yours. Predictions were done very similar to yours, and it is related to Frank Gifford's paper this morning. Using the lognormal distribution, the predictions worked very well as long as you stayed near the mean. But when you went to the extremes, if you tried to predict the extremes near ninety-nine percent, which is needed for many problems, the whole system fell apart. While near the means it worked very well.

Benarie: It's quite a general statistical property that if you don't have any infinite samples, then at the ends of the sample distribution you go wrong. That's sure. One more point, I stressed that I am interested here in long-time means, and as you have seen in the first table, I neglected the stable and the unstable situations, saying that the mean is mainly influenced by the 75% of neutral situations. I know that I can't do anything with the extremes.

Singer: It's true, but I know the normal situation. People will then take your curve and extrapolate it to 99 percent.

Benarie: No, they shouldn't do that.

Arnold Court: The apparent relation between the distributions of the concentrations and the wind speeds may be valid, but this does not mean that either is necessarily lognormal. For one thing, we as meteorologists and climatologists cannot accept lognormality for wind speeds. By the argument which the speaker made earlier in the discussion, we are looking for the most simple relation. Winds are basically three-dimensional vectors. The third dimension, up, is generally one or two orders of magnitude less than the two horizontal vectors, so we tend to ignore it. However our general attitude toward wind is that we have two orthogonal components, and we represent it by a bivariate distribution. We generally accept the bivariate normal largely in default of any other bivariate distribution that we can handle. If winds and components are bivariate normal, then the wind speed itself, independent of direction, has a Chi distribution of two degrees of freedom, also called the Rayleigh distribution. Now this is quite similar in appearance to a lognormal, but is a different distribution. On the other hand if you accept lognormality for wind speed, you have a very difficult time deriving the distribution for winds by components. Therefore I think that if the speaker's argument holds that the distributions of concentrations and wind speeds must be similar, this indicates that concentrations also may have a Rayleigh rather than a lognormal distribution.

Benarie: Thank you very much and mostly I agree with you, Mr. Court. Firstly, I stressed one of your points in my appendix which I didn't read here. Normally it's known that wind speed having lots of zeros is not a function to be represented by a lognormal. I am asking the meteorologists present here if they can provide me any data. I have made some experiments with a sensitive thermistor anemometer in a wind field. Because it's not a cup anemometer, it registers lots of values down near zero. It seems there are values everywhere. As I told already at the end of Mr. Gifford's paper, I am looking for a convenient engineering fit and an easy mathematical manipulation and not a theoretical explanation. Lognormal is good for me, but the argument is open as to how far it's physically good, and I leave it open.

Joseph Knox: I would like to ask you a question if I may about two of the figures pertaining to direction 108° to 121° in regard to pollutants SO_2 and NO_2. These figures have different slopes for the pollutants on lognormal paper, and the wind is shown as being approximately a lognormal function paralleling the NO_2 distribution. Since the slopes for these two pollutants are different, it doesn't parallel the SO_2 distribution.

Benarie: It should not be. I stressed in the paper that for the NO_2, which is a non-thermal emission, parallelity is requested. For a thermal emission, if we put concentration against distance with the parameter of wind speed, by a combination of the effective stack height with a formula like Brigg's, and a diffusion formula, you get . . . . (writing on board) things like that. In this case the concentrations have a slope in logarithmical representation. The exponent is -1 only when there is no thermal elevation. With thermal elevation it is different from 1. But I stress the point that you can choose a chimney thermal elevation formula which fits just the numerical value which gives you a good slope.

Knox: My point is this, I would also be afraid of chemical reactivity or photochemical reactivity affecting these distributions. As noted by Larsen several years ago, the pollutants with the steepest slopes for the largest standard geometrical deviations on lognormal paper are the most reactive pollutants. And so, I really want to comment that I see some cause for caution about dealing with photochemical reactive pollutants in this manner.

Benarie: It has to be remembered; your argument is quite valid. I asserted up to now that only the thermal rise is a cause of the variation of this ratio from 1. Another could be a sink, a reaction. Because I cannot yet give a quantitative evaluation or a theoretical explanation of the differing values, I note them only, so any tenative explanation is good.

6. AVERAGING TIME AND MAXIMA FOR DEPENDENT OBSERVATIONS

RICHARD E. BARLOW

Department of Industrial Engineering and Operations Research
and Department of Statistics
University of California, Berkeley, California

and

NOZER D. SINGPURWALLA

Department of Operations Research
The George Washington University, Washington, District of Columbia.

Introduction

Monitoring Air Pollutant Concentrations

Under the Continuous Air Monitoring Program of the Environmental Protection Agency, pollutant concentrations are punched into a computer tape every five minutes. Let $t_1, t_2, \ldots, t_k, \ldots, t_n$ denote the instants of time, spaced five minutes apart, at which concentrations of a certain pollutant, say $x_{t_1}, x_{t_2}, \ldots, x_{t_k}, \ldots, x_{t_n}$ are recorded on the tape (Larsen (1969)).

We assume, for now, that the observations represent a time series in which the successive observations are highly correlated. Consider averages of length k

$$k^{-1}\left[x_{t_1} + x_{t_2} + \cdots + x_{t_k}\right],\ k^{-1}\left[x_{t_{k+1}} + x_{t_{k+2}} + \cdots + x_{t_{2k}}\right],$$

$$\cdots\ k^{-1}\left[x_{t_{n-k+1}} + \cdots + x_{t_n}\right]$$

where $k << n$.

For purposes of evaluating air quality, it is important to know the probability of maximum pollutant concentrations exceeding state standards which are stated for various averaging times. Let

$$\eta_{k,n} = \mathrm{Max}\left\{k^{-1}\left[x_{t_1} + \cdots + x_{t_k}\right], \right.$$
$$\left. \ldots, k^{-1}\left[x_{t_{n-k+1}} + \cdots + x_{t_n}\right]\right\}$$

We are interested in obtaining the distribution of $\eta_{k,n}$ for k moderate and n large.

A Survey of Results Assuming Independence

If the sequence of observations x_{t_i}, $i = 1,2, \ldots, n$ were independent, as was assumed by Barlow (1972) and Singpurwalla (1972), we could use extreme value theory to determine the limiting distribution of $\eta_{k,n}$ as a function of the averaging time k. Under the hypothesis of independence, it is easy to verify that when the distribution of pollutant concentration, F, is assumed to be either a normal, a lognormal, a gamma or a Weibull

$$\lim_{n\to\infty} P\left[\frac{\eta_{k,n} - \beta_{k,n}}{\alpha_{k,n}} \le x\right] = \exp(-e^{-x}) = \Lambda(x) \qquad (1)$$

$$-\infty < x < \infty$$

exists and is nondegenerate, where $\alpha_{k,n} > 0$ and $\beta_{k,n}$ are a sequence of norming constants.

Let $G(x) = 1 - e^{-x}$ for $x \geqslant 0$ and

$$R_k(x) = G^{-1}F_k(x)$$

where F_k is the k-fold convolution of F with itself. Gnedenko (1943) (cf. Marcus and Pinsky (1969)) showed that the norming constants could be expressed as

$$\beta_{k,n} = \frac{R_k^{-1}(\log n)}{k} \qquad (2)$$

and

$$\alpha_{k,n} = \frac{R_k^{-1}(1 + \log n) - R_k^{-1}(\log n)}{k} \qquad (3)$$

Hence for large n,

$$P(\eta_{k,n} \leq x) \sim \Lambda\left[\frac{x-\beta_{k,n}}{a_{k,n}}\right] \tag{4}$$

$\beta_{k,n}$ is the location parameter and also approximately the 37th percentile of

$$\Lambda\left[\frac{x-\beta_{k,n}}{a_{k,n}}\right]$$

and thus provides a convenient way of summarizing $\eta_{k,n}$.

The main difficulty in using $\beta_{k,n}$ occurs in computing the convolution F_k. In the case where F is the gamma (normal) distribution, then of course F_k is again a gamma (normal) distribution and there is no problem in computing $R_k(x)$.

For large n, and $k << n$, Gurland (1955) has approximated $\beta_{k,n}$ and $a_{k,n}$ when F is the gamma distribution, i.e.,

$$F(x) = \int_0^x \frac{u^{\lambda-1} e^{-u/\theta}}{\theta^{\lambda}\Gamma(\lambda)} du$$

For this case

$$\beta_{k,n} \sim \frac{\theta}{k}(\log n), \text{ and } a_{k,n} \sim \frac{\theta}{k} \tag{5}$$

If we let

$$F(x) = \Phi\left[\frac{x-\mu}{\sigma}\right] \qquad -\infty < x < \infty$$

where

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

that is, if F is a normal distribution with mean μ and variance σ^2, then we can immediately verify (cf. Cramer (1951) pp. 374-375) that for large n

$$\beta_{k,n} \sim \sigma k^{-1/2} (2 \log n)^{1/2} + \mu, \tag{6}$$

$$\text{and } a_{k,n} \sim \sigma k^{-1/2} (2 \log n)^{-1/2}$$

Barlow (1972) Corollary 4.3 has obtained bounds on $\beta_{k,n}$ when F is continuous, $F(0) = 0$, and $R(x)$ is convex (concave). He has shown that

$$\frac{1}{k} R^{-1} \Gamma_k^{-1} G(\log n) \le (\ge) \beta_{k,n} \le (\ge) R^{-1}\left[\frac{1}{k} \Gamma_k^{-1} G(\log n)\right] \tag{7}$$

where

$$\Gamma_k(x) = 1 - e^{-x}\left[\sum_{j=0}^{k-1} \frac{x^j}{j!}\right] \quad \text{for } x \ge 0$$

is the gamma distribution and $G \equiv \Gamma_1$.

For example, if we let

$$F(x) = 1 - \exp\left\{-\left[\frac{x}{\delta}\right]^{1/b}\right\} \quad x \ge 0$$

that is, if F is a Weibull distribution with scale parameter δ and shape parameter $1/b$, then for large n Equation 7 gives

$$k^{-1} \delta (\log n)^b \le \beta_{k,n} \le k^{-b} \delta (\log n)^b \tag{8}$$

when $0 < b < 1$.

Motivation and Summary

Since air pollutant data are often correlated, as will be illustrated in the next section, the assumption of independence for the sequence x_{t_i} , $i = 1,2, \ldots, n$ is clearly incorrect. We can overcome this difficulty if it is reasonable to assume that the sequence of observations $[x_{t_i}]$ is *associated*. Association is a strengthening of the concept of positive correlation and is defined and discussed in the next section. In that section we show that certain air pollutant data can be modelled by an autoregressive process of suitable order. In the next section and the one following it, we show that the extreme value approximation function given by Equation 4 is a lower bound on the distribution function of the maxima of averages of associated observations. Based on this result, $\beta_{k,n}$, (or its upper bound), is an upper bound on the 37th percentile of the distribution of the maxima of averages of associated observations.

Preliminaries

Suppose than n observations

$$x_{t_1}, x_{t_2}, \ldots, x_{t_k}, \ldots, x_{t_n}$$

which are generated sequentially in time represent a *discrete time series.* We regard these observations as a particular realization of a stochastic process.

We focus attention on those processes which are *strictly stationary.* For such processes, the joint distribution of any set of observations is unaffected by shifting all the times of observation forward or backward by any integer amount k. The mean μ of the process can be estimated as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_{t_i}$$

and the variance σ_x^2 of the process can be estimated as

$$\hat{\sigma}_x^2 = \frac{1}{n} \sum_{i=1}^{n} \left[x_{t_i} - \bar{x} \right]^2$$

The covariance between X_{t_i} and $X_{t_{i+k}}$ is called the *autocovariance* at lag k, and is defined as (capital X_{t_i}'s are random variables)

$$\gamma_k = \text{Cov}\left[X_{t_i}, X_{t_{i+k}} \right] = E\left[(X_{t_i} - \mu)(X_{t_{i+k}} - \mu) \right]$$

For a stationary process, the *autocorrelation* at lag k is defined as

$$\rho_k = \frac{\gamma_k}{\gamma_o}$$

The most satisfactory estimate of ρ_k is given as (cf. Box and Jenkins (1969)).

$$r_k = \frac{C_k}{C_o} \qquad (9)$$

where

$$C_k = \frac{1}{n} \sum_{i=1}^{n-k} \left[x_{t_i} - \bar{x} \right] \left[x_{t_{i+k}} - \bar{x} \right], \; k = 0, 1, 2, \ldots, K.$$

In practice, to obtain a useful estimate of the autocorrelation function, we would need at least 50 observations and the estimated autocorrelations r_k would be calculated for k = 0,1,2, . . ., K, where K is not larger than n/4. (cf. Box and Jenkins (1969), p. 33.)

Associated Processes and Air Pollutant Measurements

Random variables $X_1, X_2, \ldots, X_n$ are said to be *associated* if

$$\text{Cov}\,(\Gamma[\underline{X}], \Delta[\underline{X}]) \geq 0$$

for all pairs of *binary, increasing* functions Γ and Δ where

$$\underline{X} = (X_1, X_2, \ldots, X_n)$$

(Binary functions are 0 and 1 valued functions.) Essentially, this is a strengthening of the concept of nonnegative correlation. The definition is due to Esary, Proschan and Walkup (1967), who also prove many important properties of associated random variables. For example, two *binary* random variables X and Y, are associated if and only if

$$\text{Cov}\,(X,Y) \geq 0$$

This is *not* true for arbitrary random variables. They also show that *independent* random variables are associated.

It follows easily from the definition that *increasing functions* (not necessarily binary) of associated random variables are associated. Hence if air pollutant measurements

$$X_{t_1}, X_{t_2}, \ldots, X_{t_n}$$

are associated then so are their averages

$$\frac{1}{k}\left[X_{t_1} + \cdots + X_{t_k}\right], \frac{1}{k}\left[X_{t_{k+1}} + \cdots + X_{t_{2k}}\right],$$

$$\ldots, \frac{1}{k}\left[X_{t_{n-k+1}} + \cdots + X_{t_n}\right]$$

Now let $[X_\tau ; \tau \in D]$ be a stochastic process, where $D = [1,2,3, \ldots]$ or $D = [0,\infty]$, for example. The process is said to be *associated* if, for all $[\tau_1, \tau_2, \ldots, \tau_n] \in D$ (the τ_i's need not be equally spaced) and all $n \geq 1$, the random variables

$$X_{\tau_1}, X_{\tau_2}, \ldots, X_{\tau_n}$$

are associated. The definition can be found in Esary and Proschan (1970). They study special performance processes of interest in reliability theory. It follows from the definition of an associated process that the autocorrelation function, $\rho(t) \geq 0$. However, $\rho(t) \geq 0$, does not of course imply in general that the process is associated. Additional restrictions on $\rho(t)$ are required, in general, to assure association.

Air pollutant concentrations follow a diurnal cycle which results in an autocorrelation which may assume both positive and negative values. Hence we cannot expect hourly averages to be associated. If we record only the high-hour daily average the association concept is more reasonable if we also confine observation to a single season. Figure 1 is a plot of oxidant data for Livermore, California covering the period June-August 1970. The sample autocorrelation shown in Figure 2 shows the existence of a 6 - 8 day weather cycle phenomenon.

Since the autocorrelation shows negative values, it is unreasonable to assume oxidant values are *associated* in time according to our definition. However, oxidant is a secondary pollutant and highly dependent on meteorological conditions. Figures 3 and 4 are plots of carbon monoxide data for Livermore, California covering the period June-August, 1970. The autocorrelation seems to remain positive within the range of sampling error. The assumption of association may be reasonable for primary pollutants over a time period not exceeding a season. Also, the less dependent the pollutant is on the weather cycle, the more likely the assumption of association will be valid. As we shall see, the association assumption, when valid, will enable us to obtain useful bounds on quantities of interest.

Ash, Bloomfield and McNeil (1972) have used a fourth root transformation on SO_2 data. The resulting data was modelled using a *Brownian motion* process. Such processes have independent increments and are always associated, since independent random variables are associated.

The Autoregressive Process

Most of the time series occuring in practice can be reasonably well explained by an autoregressive process. In this section, abstracted from Box and Jenkins (1969), we review some well-known properties of such processes.

The models that are usually employed in time series analysis are based on the idea that a time series in which the successive values $X_{t_1}, X_{t_2}, \ldots$ are highly dependent, can be regarded as generated from a series of independent shocks $[a_{t_i}]$ i = 1,2, ... These shocks are random drawings from a fixed distribution, usually assumed normal, and having a mean zero and a variance σ_a^2. The $[a_{t_i}]$ process is transformed to the $[X_{t_i}]$ process by what is known as a *linear filter.*

A $[X_{t_i}]$ process which is extremely useful for representing certain practically occuring situations is called the *autoregressive* process. Let $X_{t_i} - \mu = \tilde{X}_{t_i} = 1,2, \ldots$. Then the process

$$\tilde{X}_{t_n} = \phi_1 \tilde{X}_{t_{n-1}} + \phi_2 \tilde{X}_{t_{n-2}} + \cdots + \phi_p \tilde{X}_{t_{n-p}} + a_{t_n}$$

is called an autoregressive process of order p. In the next section we establish conditions under which an autoregressive process is associated.

If we define a *backward shift operator* B as

$$BX_{t_n} = X_{t_{n-1}}$$

then, the above autoregressive process can be written as

$$\phi(B)\tilde{X}_{t_n} = a_{t_n}$$

where $\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)$.

The equation $\phi(B) = 0$ is called the *characteristic equation* of the process.

Several properties of the autoregressive process have been given by Box and Jenkins (1969). We summarize below a few pertinent ones.

(a) An autoregressive process is stationary if the roots of its characteristic equation lie outside the unit circle.

(b) The autocorrelation function ρ_k of an autoregressive process satisfies a difference equation whose general solution is

$$\rho_k = A_1 G_1^k + A_2 G_2^k + \cdots + A_p G_p^k$$

where G_i^{-1} are the roots of the characteristic equation. Thus, the autocorrelation function of an autoregressive process tails off either exponentially, or as a mixture of exponentials and damped sine waves, depending on the nature of the roots G_i^{-1} (or equivalently, the parameters ϕ_i).

(c) If we let

$$\phi = (\phi_1, \phi_2, \ldots, \phi_p), \quad R_p = (\rho_1, \rho_2, \ldots, \rho_p),$$

and

$$\pi_p = \begin{vmatrix} 1 & \rho_1 & \rho_2 \cdots \rho_{p-1} \\ \rho_1 & 1 & \rho_1 \cdots \rho_{p-2} \\ \vdots & & \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} \cdots 1 \end{vmatrix}$$

then $\phi = \pi_p^{-1} R_p$ can be used to obtain what are known as the *Yule Walker* estimates of the parameters ϕ_i, by replacing the ρ_i by their estimates r_i.

(d) The *partial autocorrelation* function of an autoregressive process is defined as

$$\phi_{kk} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_2 \cdots \rho_1 \\ \rho_1 & 1 & \rho_1 \cdots \rho_2 \\ \vdots & & \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} \cdots \rho_k \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 \cdots \rho \\ \rho_1 & 1 & \rho_1 \cdots \rho \\ \vdots & & \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} \cdots \rho_k \end{vmatrix}} \tag{10}$$

For an autoregressive process of order p , the partial autocorrelation function ϕ_{kk} will be non-zero for k less than or equal to p, and will be zero for k greater than p.

Estimates for the partial autocorrelation function can be obtained by using r_k in place of ρ_k. If $\hat{\phi}_{kk}$ is an estimator of ϕ_{kk} then

$$\mathrm{Var}\ (\hat{\phi}_{kk}) \simeq \frac{1}{n} \qquad k \geq p+1 ,$$

and this can be used to test if the partial autocorrelation function has a cut-off at lag (p + 1).

Examples

As an example, we consider an autoregressive process of order two. Most of the time series commonly occuring in practice can be described by this process. The process can be written as

$$\bar{X}_{t_n} = \phi_1 \tilde{X}_{t_{n-1}} + \phi_2 \tilde{X}_{t_{n-2}} + a_{t_n}$$

For stationarity, the roots of $1-\phi_1 B-\phi_2 B^2 = 0$ must be outside the unit circle. This implies that the parameters ϕ_1 and ϕ_2 must lie in the triangular region given by

$$\phi_1 + \phi_2 < 1 \tag{11}$$

$$\phi_2 - \phi_1 < 1$$

$$-1 < \phi_2 < 1$$

If G_1^{-1} and G_2^{-1} are the roots of the characteristic equation, the autocorrelation function is

$$\rho_k = \frac{G_1\left[1-G_2^2\right] G_1^k - G_2\left[1-G_1^2\right] G_2^k}{\left[G_1-G_2\right]\left[1+G_1 G_2\right]}$$

When the roots are real (i.e., $\phi_1^2 + 4\phi_2 > 0$), the autocorrelation function consists of a mixture of damped exponentials. Additionally, if ϕ_1 and ϕ_2 are both positive, the process is associated and the autocorrelation function remains positive as it damps out. If the roots are complex the autocorrelation function damps out sinusoidally. A necessary condition for the association of an autoregressive process of order two with positive coefficients, is that its autocorrelation function remain positive as it damps out. The coefficients ϕ_1 and ϕ_2 can be estimated using the relationships

$$\phi_1 = \rho_1(1-\rho_2)/(1-\rho_1^2) \tag{12}$$

$$\phi_2 = (\rho_2-\rho_1^2)/(1-\rho_1^2)$$

Application to Carbon Monoxide Data

We estimate the autocorrelation function ρ_k of the carbon monoxide data given in Figure 3 using Equation 9. The estimates are $r_1 = .736$, $r_2 = .676$, $r_3 = .560$, $r_4 = .461, \ldots$. The partial autocorrelation function $\phi_{k,k}$ is estimated using Equation 10 and replacing the ρ_i's by the r_i's, for $k = 2$ and $k = 3$. These estimates are $\hat{\phi}_{22} = .294$ and $\hat{\phi}_{33} = -.018$. Since $\hat{\phi}_{33} \approx 0$, it is reasonable to conclude that the carbon monoxide data can be reasonably well described by an autoregressive process of order 2.

Estimators of the parameters of the autoregressive process ϕ_1 and ϕ_2 are obtained using Equation 12 and by replacing ρ_1 and ρ_2 by their estimates r_1 and r_2 respectively. These estimates are $\hat{\phi}_1 = .520$ and $\hat{\phi}_2 = .293$. In the next section we shall show that this process is also associated.

Associated Stochastic Processes

We now establish conditions on various stochastic processes which ensure association. The concept of association is then used to establish probability bounds on the distribution of the maxima of averaging times.

Associated Autoregressive Processes

Let

$$X_{t_n} = \sum_{i=1}^{p} \phi_i X_{t_{n-i}} + a_{t_n}$$

be the pth order autoregressive process discussed in the previous section, entitled "The Autoregressive Process". Here we only assume that the a_{t_i}'s are independent and identically distributed. The process need *not* be stationary. If the process is associated then it follows that, when $p = 1$,

$$\text{Cov}\,(X_{t_1}, X_{t_2}) = (\text{Var}\,X_{t_1})\,\phi_1 \geq 0$$

which implies $\phi > 0$.

Lemma 1:

If $\phi_i \geqslant 0$ $(i = 1,2, \ldots, p)$, then the autoregressive process of order p is associated. (See also Theorem 2.)

Proof:

Esary, Proschan and Walkup (1967) prove that independent random variables are associated and also that increasing functions of associated random variables are associated. Hence

$$X_{t_1} = a_{t_1}$$

$$X_{t_2} = \phi_1 a_{t_1} + a_{t_2}$$

and

$$X_{t_3} = (\phi_1^2 + \phi_2)\, a_{t_1} + \phi_1 a_{t_2} + a_{t_3}$$

are associated. The lemma follows by induction. ||

Clearly, it follows from the lemma and the previous remarks that an autoregressive process of order 1 is associated if and only if $\phi_1 \geqslant 0$.

Bounds on the Distribution of the Maxima of Averages for Stationary Associated Processes

Let $X_{t_1}, X_{t_2}, \ldots, X_{t_k}, \ldots, X_{t_n}$ be an associated process and

$$\eta_{k,n} = \mathrm{Max}\left\{ k^{-1}\left[X_{t_1} + \cdots + X_{t_k}\right], k^{-1}\left[X_{t_{k+1}} + \cdots + X_{t_{2k}}\right], \ldots, k^{-1}\left[X_{t_{n-k+1}} + \cdots + X_{t_n}\right]\right\}.$$

Lemma 2:

Let $[X_{t_1}, X_{t_2}, \ldots]$ be a stationary associated process with marginal distribution F, $\beta_{k,n}$ and $a_{k,n}$ as defined in the introductory section. If F_k is such that Equation 1 holds, then

$$P\left[\frac{\eta_{k,n} - \beta_{k,n}}{a_{k,n}} \leq x\right] \geq \exp\left(-e^{-x}\right)$$

for sufficiently large n.

Proof:

Let $X_1, X_2, \ldots, X_n$ be associated random variables. Esary, Proschan and Walkup (1967, pp. 1472-73) prove that

$$P\left[\mathrm{Max}\,(X_1, X_2, \ldots, X_n) \leq x\right] \geq \prod_{i=1}^{n} P(X_i \leq x) \qquad (13)$$

where the right-hand side corresponds to the case of *independent* random variables. In the introductory section we noted that, for n large

$$\prod_{i=0}^{\frac{n}{R}-1} P\left[(\xi_i - \beta_{k,n})/\alpha_{k,n} = x\right] \sim \Lambda(X)$$

where,

$$\xi_i = \frac{X_{t_{ik+1}} + \cdots + X_{t_{ik+k}}}{k}$$

so that the lemma follows from Equation 13.||

Example

Let $[X_t;\, t \epsilon D]$ be a pth order Gaussian autoregressive process such that $\phi_i > 0$ for $i = 1, 2, \ldots, p$. Then from Equation 6 and the previous lemmas it is associated and (assuming $\mu = 0$)

$$P\left[\frac{\eta_{k,n} - k^{-1/2}\sigma\,(2\log n)^{1/2}}{\sigma\, k^{-1/2}\,(2\log n)^{-1/2}}\right] \geq \Lambda(x) \qquad (14)$$

for large n, where σ is the standard deviation of the process. The extreme value approximation is not useful, in general, for bounding the tails of the distribution of $\eta_{k,n}$. (See Cramér (1951) p. 377.)

Application to Carbon Monoxide Data

In the section on the autoregressive process we showed that the carbon monoxide data given in Figure 3 can be reasonably well described by a stationary autoregressive process of order 2 with positive coefficients ϕ_1 and ϕ_2. It therefore follows from Lemma 1 that the process is also associated. If we next assume that the independent random shocks $[a_{t_i}]$, $i = 1, 2, \ldots$, discussed in the section on the autoregressive process have a Gaussian distribution, then the conditions of this example apply.

The California state standard for carbon monoxide is specified at 20 ppm for an 8-hour averaging time. The carbon monoxide data for Livermore, California during the months of June, July and August, 1970 reveals that the above standard was never violated. Using this data (presented in Figure 3), we would like to compute a lower bound on the probability that the specified State standard will be violated. The standard deviation of this data, σ, was estimated as 2.8. Basing our total sample size as hourly observations for 90 days, we have $n = 2160$, and considering averages of length 8 (because of the 8-hour averaging time), we take $k = 8$. Thus

$$P\left[\eta_{k,n} \le 20\right] \ge \Lambda\left[\frac{20-8^{-1/2}(2.8)(2\log 2160)^{1/2}-5.16}{8^{-1/2}(2.8)(2\log 2160)^{-1/2}}\right]$$

$$= \Lambda\,(43.39)$$

$$\therefore\ P(\eta_{k,n} \le 20) \ge e^{-e^{-(43.39)}} \approx 1$$

and this bears out the fact there were no violations of the specified standard.

In the light of the observed data it appears that the specified standard is unreasonably high.

In general, if F is difficult to convolute and if $R(x) = -\log\,[1 - F(x)]$ is convex, then

$$\beta_{k,n} \le R^{-1}\left[\frac{1}{k}\Gamma_k^{-1}\,G(\log n)\right]$$

as noted in Equation 7. Let $\xi_{.37}$ be the 37th percentile of $P[\eta_{k,n} < x]$. Thus, even in the presence of association

$$\xi_{.37} \le \beta_{k,n} \le R^{-1}\left[\frac{1}{k}\Gamma_k^{-1}\,G(\log n)\right]$$

Additional Associated Processes

It is difficult, in general, to verify that a process is associated from the definition of association. Another useful concept which implies association is that of *conditionally increasing in sequence.*

Definition:

Random variables $X_1, X_2, \ldots, X_n$ are *conditionally increasing in sequence* if

$$P(X_i > x \mid X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1})$$

is increasing in $x_1, x_2, \ldots, x_{i-1}$ for $i = 1,2, \ldots, n$.

A stochastic process is conditionally increasing in sequence if any subset of random variables based on the process is conditionally increasing in sequence.

This concept is due to Esary and Proschan (1968) who also proved the following theorem.

Theorem 1: (Esary and Proschan)

If $X_1, X_2, \ldots, X_n$ are conditionally increasing in sequence, then $X_1, X_2, \ldots, X_n$ are associated.

The concept has an immediate application to autoregressive processes of order p, which, according to Theorem 2, are associated if $\phi_i \geq 0$, for $i = 1, 2, \ldots, p$.

Theorem 2:

Autoregressive processes of order p are conditionally increasing in sequence if and only if $\phi_i \geq 0$, $i = 1.2, \ldots, p$.

Proof:

For an autoregressive process of order p, it is easy to verify that

$$P\left[X_{t_n} > x \mid X_{t_{n-1}} = x_{t_{n-1}}, \ldots, X_{t_{n-p-1}} = x_{t_{n-p-1}}\right]$$

is increasing in $x_{t_{n-1}}, \ldots, x_{t_{n-p-1}}$ if and only if $\phi_{x_i} \geq 0$ for $i = 1, 2, \ldots, p$.||

Lemma 3:

If $[X_t; t \epsilon D]$ is a Markov process and if

$$P(X_t > x \mid X_s = y)$$

is increasing in y for $s \leq t$, then the process is associated.

Proof:

It is sufficient to prove that the process is conditionally increasing in sequence, i.e.,

$$P\left[X_{t_n} > x \mid X_{t_1} = x_{t_1}, \ldots, X_{t_{n-1}} = x_{t_{n-1}}\right]$$

But the Markov property implies that this equals

$$P\left[X_{t_n} > x \mid X_{t_{n-1}} = x_{t_{n-1}}\right]$$

which completes the proof. ||

Theorem 3:

A stationary, Gaussian process with autocorrelation function

$$\rho(t) = \int_0^\infty e^{-\lambda t}\, dH(\lambda)$$

for some distribution, H, on $[0,\infty]$ is associated. (Note that time may be either continuous or discrete.)

Proof:

It is well known that a stationary Gaussian process is completely determined by its autocorrelation function together with the marginal mean and variance. By Lemma 3, the stationary Gaussian *Markov* process with autocorrelation

$$\rho(t) = e^{-\lambda t}$$

is associated

(Note $P[X_t > x | X_s = y) = \int_{x-\rho y} \exp[-u^2/[2(1-p^2)]]\, du/(2\pi(1-\rho^2))^{1/2}$.)

To complete the the proof, let $p_i \geq 0$, $i = 1,2, \ldots, k$ and $\sum_{i=1}^{k} p_i = 1$. Also specify $\lambda_i > 0$, $i = 1,2, \ldots, k$. Let $[X_i(t); t \geq 0]$ be a stationary Gaussian process with autocorrelation

$$\rho(t) = e^{-\lambda_i t}$$

for $i = 1,2, \ldots, k$. Assume that the k processes are mutually independent. Since each process is associated, it follows that the process

$$Y_t = \sum_{i=1}^{k} \sqrt{p_i}\, X_i(t)$$

is associated. (Recall that increasing functions of associated random variables are associated.) Also

$$\rho(s) = \operatorname{Cov}\left[Y(t),\ Y(t+s)\right]$$

$$= \operatorname{Cov}\left[\sum_{i=1}^{k} \sqrt{p_i}\, X_i(t),\ \sum_{i=1}^{k} \sqrt{p_i}\, X_i(t+s)\right]$$

$$= \sum_{i=1}^{k} p_i e^{-\lambda_i t}$$

By a limiting argument we can show that if

$$\rho(t) = \int_0^\infty e^{-\lambda t}\, dH(\lambda)$$

and the process is a stationary Gaussian process then the process is associated. ||

The previous theorem has useful applications to data which is believed to be generated by a stationary Gaussian process. If we can approximate the sample autocorrelation function by a convex combination of exponentials then this is evidence that the process is associated.

Discrete State Markov Processes

An example of a discrete state Markov process is the birth and death process assuming states [0,1,2, . . .]. Such processes, it turns out, are always associated if the time variable $t \geq 0$, can assume any non-negative value. When such processes are restricted to integer values they of course remain associated. On the other hand, a random walk process in discrete time with transition matrix

$$\begin{bmatrix} b & c & 0 & 0 & \cdots & \\ a & b & c & 0 & 0 & \\ 0 & a & b & c & 0 & \cdots \\ \cdots & & & & & \end{bmatrix}$$

is associated if and only if $b^2 \geq a\,c$.

The above remarks follow from

Theorem 4:

If $[X_t;\, t \in D]$ is a Markov process with transition probability matrix $(P_{ij}(t))$ which is *totally positive* in i and j for all $t \geq 0$; i.e.,

$$\begin{vmatrix} P_{i_1,j_1}(t) & P_{i_1,j_2}(t) \\ P_{i_2,j_1}(t) & P_{i_2,j_2}(t) \end{vmatrix} \geq 0$$

then the process is associated. (The time variable may be either continuous or discrete.) We assume here that $i_1 \leq i_2$ and $j_1 \leq j_2$.

Theorem 4 was proved by D. J. Daley (1968).

Karlin (1968) showed that birth and death processes with state space [0,1,2, . . .] always satisfy the conditions of Theorem 4 and hence are associated. Esary and Proschan (1970) showed that two-state birth and death processes are associated.

Conclusion

Our objective in this paper has been to present a new and different approach to the analysis of air pollution data, which can be, and perhaps should be modelled as a time series. The results presented here are based on more realistic considerations than those of a similar nature presented before, and should be useful in setting and monitoring air pollution standards.

Though the primary motivation in this paper has been the analysis of air pollution data, the results obtained here should be of a more general interest. The results on associated stochastic processes presented in the section headed "Associated Stochastic Processes" should have applications in time series analysis, queueing theory, and reliability theory.

By showing that the extreme value distribution is a lower bound on the distribution function of the maxima of observations generated by an associated stochastic process, we have expanded the scope of applications of extreme value theory. However, the extreme value approximation may be too conservative in many applications.

Theorem 3 asserts that if $[X_t;\ t > 0]$ is a stationary Gaussian process and $\rho(t)$ can be represented as a mixture of exponentials, then

$$P\left[\operatorname{Max}(X_{t_1}, X_{t_2}, \ldots, X_{t_n}) \le x\right] - \prod_{i=1}^{n} P(X_{t_i} \le x) \ge 0.$$

S. M. Berman *[Annals of Mathematical Statistics*, Vol. 35, pp. 502-516, (1964)] has shown that, in general, if $EX_{t_i} = 0$, $EX^2_{t_i} = 1$ and $EX_{t_0}X_{t_n} = r_n$, then

$$\left|P\left[\operatorname{Max}(X_{t_1}, X_{t_2}, \ldots, X_{t_n}) \le x\right] - \prod_{i=1}^{n} P\left[X_{t_i} \le x\right]\right|$$

$$\le \sum_{j=1}^{n-1} |r_j|(n-j)\phi_2(x, x; |r_j|)$$

where ϕ_2 is a two dimensional normal density with mean vector $\underline{0}$ and correlation $|r_j|$. Berman further shows that if either

$$\lim_{n \to \infty} r_n \ \log_n = 0$$

or

$$\sum_{n=1}^{\infty} r_n^2 < \infty$$

then Equation 1 holds with $\beta_{k,n}$ and $\alpha_{k,n}$ given by Equation 6 where $\mu = 0$ and $\sigma = 1$.

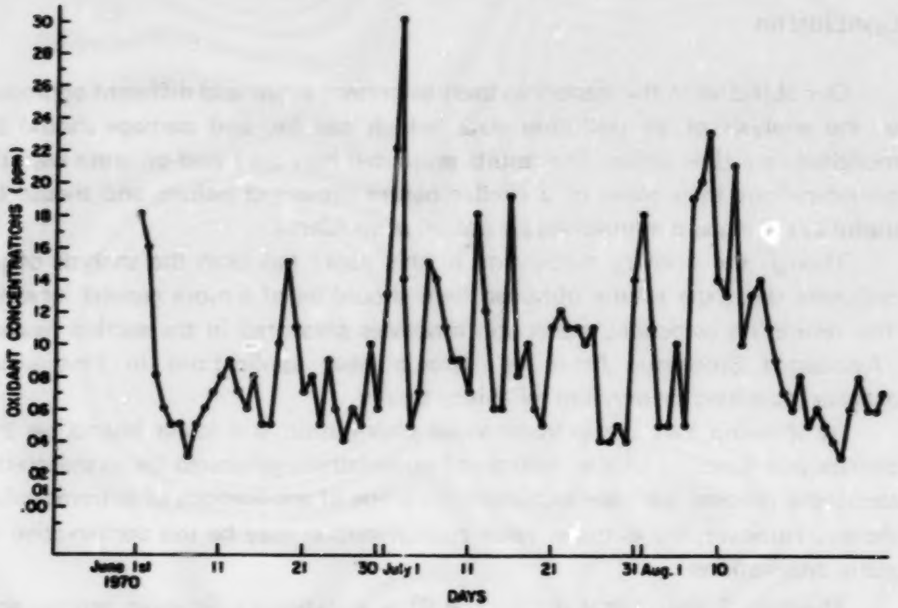


Figure 6-1. Oxidant Concentrations in ppm for Livermore, California, June - August, 1970.

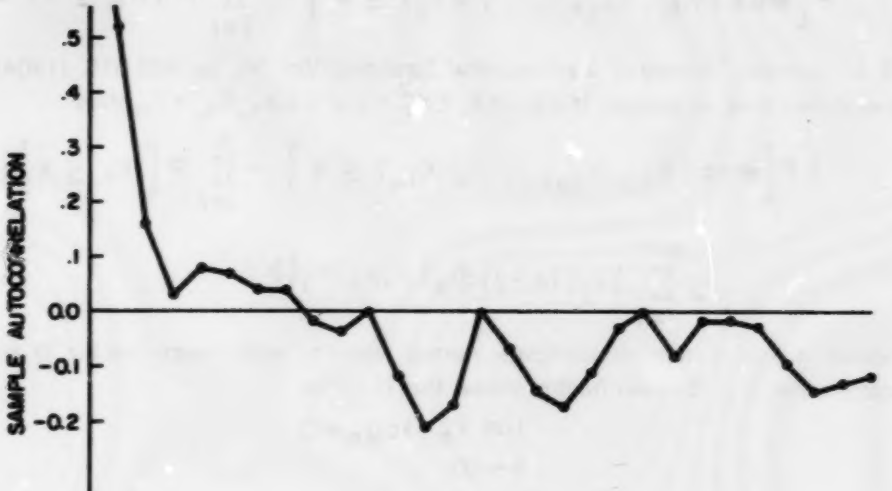


Figure 6-2. Sample Autocorrelation Function for Oxidant Data From Livermore, California, June - August, 1970.

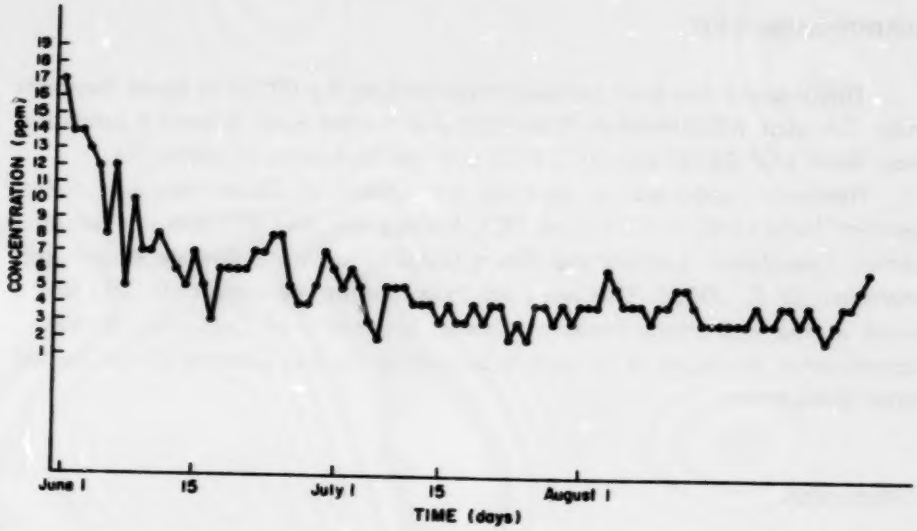


Figure 6-3. Carbon Monoxide Concentrations in ppm for Livermore, California, June - August, 1970.

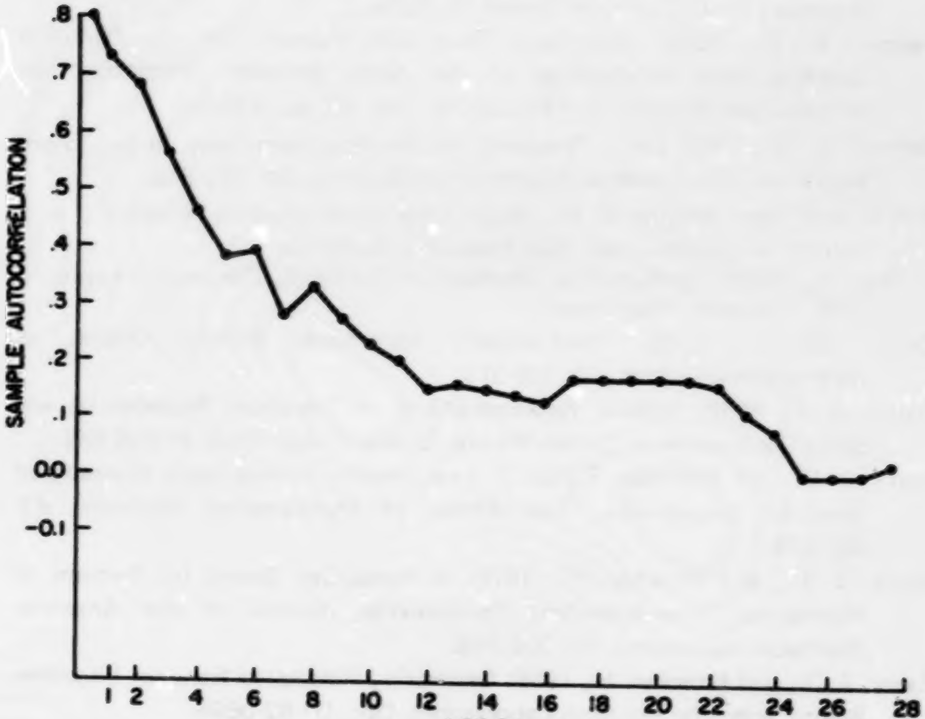


Figure 6-4. Sample Autocorrelation Function for Carbon Monoxide Data from Livermore, California, June - August, 1970.

Acknowledgement

This research has been partially supported by the Office of Naval Research under Contract N00014-69-A-0200-1036 and the National Science Foundation under Grants GP-29123 and GK-23153 with the University of California.

Research supported in part by the Office of Naval Research under Contract N00014-67-A-0214 Task 001, Project NR 347 020 and the National Science Foundation Institutional Grant GU3287 with the George Washington University, D. C. 20006. This work was begun while the author (N.D.S.) was a visitor at the Operations Research Center, University of California, Berkeley. Reproduction in whole or in part is permitted for any purpose of the United States Government.

References

Ash, D., Bloomfield, P., and McNeil, D. R., 1972: On the Statistical Analysis of Air Pollution Data. Department of Statistics, Princeton University, Princeton, N. J., Technical Report 19, Series 2.

Barlow, R. E., 1972: Averaging Time and Maxima for Air Pollution Concentrations. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. VI, pp. 433-442.

Berman. S. M., 1969: Limit Theorems for the Maximum Term in Stationary Sequences. *The Annals of Mathematical Statistics. 35:* 512-516.

Box, G. E. P., and Jenkins, G. M., 1969: *Time Series Analysis Forecasting and Control.* Holden-Day, Inc., San Francisco, California.

Cramer, H., 1951: *Mathematical Methods of Statistics.* Princeton University Press, Princeton, New Jersey.

Daley, D. J., 1968: Stochastically Monotone Markov Chains. *Z. Wahrscheinlichkeitsth. 10:* 305-317.

Daley, D. J., 1969: Integral Representations of Transition Probabilities and Serial Covariances of Certain Markov Chains. *J. App. Prob. 6:* 648-659.

Esary, J. D., and Proschan, F., 1972: Relationships Among Some Concepts of Bivariate Dependence. *The Annals of Mathematical Statistics. 43:* 651-655.

Esary, J. D., and Proschan, F., 1970: A Reliability Bound for Systems of Maintained, Interdependent Components. *Journal of the American Statistical Association. 65:* 329-338.

Esary, J. D., and Proschan F., 1968: Generating Associated Random Variables. Boeing Scientific Research Laboratories, Doc. D1-82-0696.

Esary, J. D., Proschan F., and Walkup, D. W., 1967: Association of Random Variables, With Applications. *The Annals of Mathematical Statistics. 38:* 1466-1474.

Gurland, J., 1955: Distribution of the Maximum of the Arithmetic Mean of Correlated Random Variables. *The Annals of Mathematical Statistics. 26:* 294-300.

Karlin, S., 1968: *Total Positivity, Volume I,* Stanford University Press, Stanford, California.

Larsen, R. I., 1969: A New Mathematical Model of Air Pollutant Concentration Averaging Time and Frequency. *J. Air Pollution Control Association. 19:* 24-30.

Marcus, M. and Pinsky, M., 1969: On the Domain of Attraction of $e^{-e^{-x}}$. *J. Math. Anal. Appl. 28:* 440-449.

Singpurwalla, N. D., 1972: Extreme Values from a Lognormal Law with Applications to Air Pollution Problems, Technometrics. 14: 703-711.

DISCUSSION

Don Pack: The question is as follows—in working with a time series of data you know that it is contaminated in various ways either by the position of the sampler or otherwise but I'd like to particularly direct the question toward the instrumental contamination. We have remarked that the tails of the distribution are of particular interest. Yet, if I understand instrumentation, if the dynamic range of the instrument is just about that of the range of the pollution concentrations you can expect, the maximum error will occur at the threshold and at the very high values—such things as poisoning bubblers by a spike of concentration. Can the statistician clean up the time series distribution through examination of the instrument characteristics so you don't have to examine each individual observation for its validity?

Singpurwalla: If I understand your question correctly, I would say possibly yes.

Pack: Do you know how long it takes to go through each one of these? And yet, you know that there are errors in here. Can you establish the probability that the observation is real and not instrumental?

Singpurwalla: There are techniques, outlier techniques, or there are probably techniques of some kind of pattern recognition, discriminant analysis, which could be used to put a particular piece of data in category A or category B, where category A might be something that is real, and Category B might be something that is phony for somebody else. I would imagine yes.

Pack: I simply haven't seen it done. That is why I asked.

Singpurwalla: I think it could be done.

7. A STOCHASTIC MODEL FOR ESTIMATING POLLUTANT EXPOSURE BY MEANS OF AIR QUALITY DATA

ALLAN H. MARCUS

Department of Mathematical Sciences
University of Maryland-Baltimore County
Baltimore, Maryland

Introduction

Air quality data can and should be made more useful in determining the public health implications of air pollution control strategies. The problem is that the performance of pollution control strategies is tied to time-averaged pollutant concentrations at spatially fixed sampling or monitoring stations. Different individuals in the population receive vastly different exposures from the same polluted environment. For example, an executive who drives from a nearly rural suburb to a polluted urban center business district in an enclosed air-conditioned car, and works in the upper stories of an air-conditioned office building, receives a much smaller pollutant exposure than does, say, a traffic policeman working in the same urban center. The executive may actually receive most of his dosage while waiting in a poorly ventilated parking garage for his car. An individual exposure thus depends significantly on the personal "trajectory" or movement of the individual in the urban area through space and time, and on other hard-to-predict factors. There are also important differences in individual *response* to exposure, such as the age and state of health of the person, history of smoking, and the time required for intake and elimination of pollutants by various body organs. For this reason, the urban poor (who have to live with a variety of environmental stresses, and who have a high proportion of children, elderly, ill and other susceptible types) are particularly vulnerable (U.S.E.P.A. (1971)). Age, health, income, travel patterns and other elements of "life style" are interdependent, and make the prediction of exposure and response more difficult.

The purpose of this paper is to show that many of these questions can be formulated mathematically in terms of the excursions of filtered stochastic integrals of pollutant concentration. When pollutant concentrations are functions of a Gaussian random field, some of the questions raised can be answered analytically, and most can be studied numerically.

The data base required to actually use the model for predictive purposes is enormous. These include: meteorological variables such as wind speed, height of the mixing layer, and ground-level turbulence; an inventory of the major point, line and area emission sources; demographic data for the estimation of personal trajectories for various population types; and reliable dosage-response data for various pollutants. The advantage of the analytical approach is that it may prove possible to combine much of the above data into a relatively small number of parameters which determine the level-crossing properties of the stochastic integrals. In this way it may prove possible to study simultaneously the performance of air quality standards and health effects on various segments of the population of alternative pollution control strategies, without resorting to extensive (and expensive) computer simulations.

Performance of Air Quality Standards

Air quality standards are defined in terms of average pollutant concentrations with respect to a specified averaging time T, which are not to exceed a threshold level L_T more than n_T times in a period of length S_T. We can define the air quality standards problem in the following rather formal way. Let C(R,t) be the pollutant concentration at a point R at a time t. The time-averaged concentration is

$$C_T(R,t) = \frac{1}{T} \int_{t-T}^{t} C(R,u)\, du \qquad (1)$$

Define $N_T(R,t)$ as the number of excursions of $C_T(R,t)$ above L_T during the interval of time $(t-S_T,t]$. That is, for u in the interval $(t-S_T,t]$, there will be a random number $N_T(R,t)$ of episodes in which $C_T(R,u)$ exceeds L_T continuously. Let D_j be the duration of the $j^{\underline{th}}$ such excursion, which starts at time $t_j (t-S_T<t_j<t)$. Then

$$\begin{aligned} C_T(R,u) &\geq L_T \text{ for } t_j < u < t_j + D_j \qquad (2)\\ &< L_T \text{ for } t_{j-1} + D_{j-1} < u < t_j\\ &\text{and } t_j + D_j < u < t_{j+1} \end{aligned}$$

where $j = 1, \ldots, N_T(R_1t)$.

The air quality standards requirement is thus, for a single monitoring site at R, $N_T(R,t) \leqslant n_T$ for all t and the severity of the regional air quality problem for this pollutant can be evaluated in terms of the probability of exceeding the limit during the interval of time (t_0,t_f).

$$P\left[N_T(R,t) > n_T,\ t_0 < t < t_f\right] \qquad (3)$$

Now, if there are k monitoring stations at points $R_1, \ldots, R_k$ in the region, the standards are harder to interpret. What might be meant is either that the standards are satisfied *for all* sites, so that the measure of severity of the pollution problem is

$$1 - P\left[N_T(R_1,t) \leq n_T; \ldots, N_T(R_k,t) \leq n_T;\ t_0 < t < t_f\right] \qquad (4)$$

or that the weighted average concentration

$$\bar{C}_T(t) = \sum_{i=1}^{k} w_i C_T(R_i,t) \quad \text{for } w_i > 0, \ \sum_{i=1}^{k} w_i = 1 \qquad (5)$$

not exceed L_T more than n_T times. Thus, letting $N_T(t)$ be the number of times $\bar{C}_T(u)$ exceeds L_T for $t\text{-}S_T \leq u < t$, the severity of the problem is given by

$$P\left[N_T(t) > n_T\right] \qquad (6)$$

The quantities may differ substantially.

One approach to these problems is by modeling. We could start by assuming that $C(R,t)$ is a stationary stochastic process, and develop the needed results in terms of familiar level-crossing probabilities, but this is a very difficult problem (Marcus (1972)). Even here, what we would need is the matrix of cross-correlations $\rho_{ij}(t)$ between the transformed concentration at R_i and transformed concentration t units of time later at R_j.

For purposes of evaluation of performance probabilities, it may be sufficient to consider only a two-state stochastic process

$$\begin{aligned} I_T(R,t) &= 1 \quad \text{if} \quad C_T(R,t) \geq L_T \\ &= 0 \quad \text{if} \quad C_T(R,t) < L_T \end{aligned} \qquad (7)$$

we could then enquire whether the times between successive crossings of level L_T constitute a realization of an alternating renewal process. If so, the intensity function (i.e., renewal density) and distributions of duration and frequency of exceedances are easily estimated. We could then answer the usual "quality control" question of whether or not a certain adversely high concentration proves that the *underlying* process is out of bounds.

The analysis of pollutant concentrations as a (multi-variate) time series is essential with regard to the stationarity of the underlying processes (i.e., statistical homogeneity with respect to time). We should examine concentrations at each site for:

(a) secular variations, such as trends (increase due to increasing regional population; decrease due to movement of industry or conversion to low-polluting fuels).

(b) cyclical variations, including seasonal, weekly, daily, and other regular periodic climatic or human movements.

(c) other persistent but irregular events.

Some useful first steps in the time series analysis of air pollution data have been made by Merz, et al. (1972) and by Ash, et. al. (1972). We will discuss these in more detail in a later section on "Stochastic Models for Pollutant Concentrations."

Individual Dosage Histories

What we are really interested in are the public health implications of pollutant control policies. Policies are often tied to physical performance characteristics of spatially fixed monitoring and sampling stations. These only indirectly characterize individual physiological response to pollutants. It would be more directly meaningful to measure *cumulative dosage* or *dosage-response* on an individual history basis. Let P_i be the space-time "trajectory" of person i, i.e., the entire history of his or her movements in the metro region during some interval of time. Most people travel extensively during the day and may be exposed to quite different concentrations at various times and places.

There are several possible indicators of total dosage in the interval (t_0, t_f). Total dosage is given by

$$Q(P_i) = \int_{t_0}^{t_f} \int_{P_i} C(R,u)\, du\, dR \tag{8}$$

If the threshold level L is important, then in terms of an indicator function

$$\begin{aligned} I_L(C) &= 1 \text{ if } C \geq L \\ &= 0 \text{ if } C < L \end{aligned} \tag{9}$$

we may want, for example, duration above L,

$$D_L^{(P_i)} = \int_{t_0}^{t_f} \int_{P_i} I_L\left[C(R,u)\right] du\, dR \tag{10a}$$

total dosage above L,

$$Q_L(P_i) = \int_{t_0}^{t_f} \int_{P_i} C(R,u)\, I_L\left[C(R,u)\right] du\, dR \tag{10b}$$

If there is a possible non-linear physiological response at time t due to after-effect of a pollutant contentration C(R,u) at some point T at an earlier time u, then we define an after-effect or response function f(C, t-u) so that response at t is

$$r(P_i, t) = \int_{-\infty}^{t} \int_{P_i} f\left[C(R,u), t-u\right] du \, dR \qquad (11)$$

which is a variable, stochastic response to a stochastically variable exposure for each trajectory. Because of the highly non-stationary nature of the exposures, an analytical study of the distribution of the indicators Q, Q_L, D_L or r seems less promising than a simulation study. The required individual trajectories can be estimated from demographic data.

These simulated histories could be compared with personal pollutant monitoring devices analogous to individual total radiation dosimeters. It should then be possible to more adequately evaluate epidemiological studies, e.g., data collection by the CHESS or CHAMP networks.

Stochastic Models for Pollutant Concentrations

Much of what is called "random" variation in a system is merely due to ignorance—we often do not know which factors affect the evolution of the system, or else we know (or suspect) that certain factors are significant, but cannot relate them precisely to system performance, and so choose incorrect functional relationships. If important variable factors are not included in predictions of system performance, they may contribute greatly to the unexplained "random" variation, and their exclusion could greatly modify the structure of the statistical data analysis. This is, in fact, the most serious problem in finding a stochastic model for statistical interpretation of pollutant concentration data.

The state of the art in predicting urban air pollution by multiple-source diffusion models was thoroughly explored in a symposium held here in 1969 (Stern (1970)) and the field has continued to develop rapidly. We assume the usual continuous point- and line-source Gaussian plume dispersion model. Let the ground-level monitor be affected by n_p point sources and by n_L line sources, both emitting continuously but with a possibly slowly varying emission rate. Let the mean wind speed be U m/sec. The ith point source emits Q_{P_i} micrograms/sec of a given pollutant, and is located at elevation H_i and distance S_i meters from the monitor. Let ϕ_i be the angle between the mean wind direction and the line from the source to the monitor. The downward distance is then $x_i = S_i \cos \phi_i$ and the crosswind distance is $y_i = S_i \sin \phi_i$. Similarly, let the shortest distance from the monitor to the ith infinite continuous line source be R_i, and let Q_{L_i} be its emission rate in micrograms/sec/m. if θ_i is the angle between the line source and the direction of the mean wind, then $x_i = R_i/(\sin\theta_i)$ is the downwind distance of the monitor. Then, defining the crosswind dispersion variance ϕ_y^2 by

$$\sigma_y^2 = a_y^2\, x^{2-n} \tag{12}$$

and the vertical dispersion variance σ_z^2 by

$$\sigma_z^2 = a_z^2\, x^{2-n} \tag{13}$$

where we usually have $0 < n < 1$, then

$$\begin{aligned} C(t) = {} & \sum_{i=1}^{n_p} \frac{Q_{pi}}{\pi\, a_y\, a_z\, U\, (S_i \cos\phi_i)^{2-n}} \\ & \exp\left\{-\frac{1}{2}(S_i\cos\phi_i)^n \left(\left[\frac{H_i}{a_z\, S_i \cos\phi_i}\right]^2 + \frac{\tan^2\phi_i}{a_y^2}\right)\right\} \\ & + \sum_{i=1}^{n_L} \frac{Q_{Li}(\sin\theta_i)^{n/2}}{(\pi/2)^{1/2}\, a_z\, U R_i^{\,1-(n/2)}} \\ & \exp\left[-\frac{1}{2}(R_i \csc\theta_i)^n (H_i \sin\theta_i / a_z R_i)^2\right] \end{aligned} \tag{14}$$

Thus, the pollutant concentration depends on *at least:*

(a) Fixed relative location of monitor and sources, through R_i, S_i and H_i.

(b) Wind direction as a time-varying stochastic process, through ϕ_i and θ_i.

(c) Wind speed and gustiness as stochastic processes, mainly through the products $a_y a_z U$ for point sources and $a_z U$ for line sources, and through n.

(d) Randomly variable emissions Q_{Pi} and Q_{Li}.

These factors are not independent, and will vary significantly with time and with location of the monitor.

The model (Eq. 14) involves a combination of multiplicative and additive factors, so it is not clear that a large-sample result, e.g., some version of the central limit, will yield a general functional form for the distribution of C(t) of normal or lognormal type. If the concentration is determined principally by a few strong sources, their strengths and the wind speed and direction are the major source of variation. On the other hand, if the monitor is surrounded by a large number of roughly equal sources, individual source strength and wind direction may contribute less to the distributional variation of C(t) than does wind speed and turbulence. The latter situation may be approximated in urban centers.

Some studies are available which give the distribution of wind speeds and turbulence in urban environments (Holzworth (1967); Brook (1972); Luna and Church (1972)). The wind speed distribution is significantly non-Gaussian and positively skewed (Brook (1972)), and may possibly be of lognormal form. However, wind speed and turbulence parameters are strongly dependent. Let σ_A be the elevation standard deviation in radians (note that if $n = 0$, then $a_y = \sigma_A$ and $a_z = \sigma_E$). The product $\sigma_A \sigma_E U$ (and equivalently, its reciprocal) does not have a lognormal distribution except for stability classes E and F, corresponding to low winds and night or overcast conditions (Luna and Church (1972)). These are, however, conditions of very high air pollution potential. One might suspect then, that urban air pollutant concentration distributions are a mixture of distributions, but with an approximately lognormal upper tail. This is also suggested by Fig. 7 of Holzworth (1967).

Larsen (1971) has shown that the lognormal concentration distribution is applicable to many sets of observations, such as hourly averages of SO_2 at the CAMP monitor in Washington, D. C. from 1961 to 1968. However, examination of smaller data sets shows that some are well described as lognormal, but others are not (see e.g., U.S.P.H.S. (1966)). A recent, very thorough study by Ash, Bloomfield and McNeil (1972) of CO and SO_2 concentrations in Camden and Bayonne, N. J. in 1970-1971, shows that the lognormal distribution is not particularly satisfactory, especially at low concentrations. They suggest that a "fourth-root Gaussian" distribution may be better (although suggesting a lognormal distribution for NO_2!); the fourth-root transformed concentrations also have the property of approximating a stationary Gauss-Markov (Ornstein-Uhlenbeck) random process.

What is needed is a combination of the deterministic diffusion modeling and the purely empirical statistical data analysis. The deterministic model provides a structural framework for predicting pollutant concentrations, with important variables such as wind speed and direction, and turbulence parameters, as predictors rather than unexplained sources of variation. The residual unexplained variation is the true "random" variation, and understanding its structure should suggest the most useful statistical analysis methods.

Some Useful Results on Level Crossings and Exceedances

It would be convenient to study the frequency, duration, and intervals between pollution episodes using well-known results about curve-crossings of random processes (see Cramer and Leadbetter (1967) for a rather complete exposition of known results). However, we can use the extensive body of results about Gaussian random processes only if we can first find a monotone transformation of pollutant concentrations such that the transformed concentrations approximate a realization of a Gaussian process. Explicitly, let g(C) be a monotone increasing function of C. We are looking for a representation

$$g\left[C_T(t)\right] = \sigma(T)\,Z(t) + \mu(T) \qquad (15)$$

where Z(t) is a zero-mean, unit-variance, Gaussian random process with auto-correlation function $\rho_T(t)$; the parameters may depend on the averaging time T and implicitly on the location of the monitor. Hence, if L_T is the T-hour standard, a pollution episode occurs if

$$C_T(t) \geq L_T \qquad (16)$$

i.e.,

$$Z(t) \geq h = \left[g(L_T) - \mu(T)\right]/\sigma(T) \qquad (17)$$

Now, assume that $\rho_T(t)$ has an expansion near the origin, in powers of t,

$$\rho_T(t) = 1 - \lambda_2(T)\,t^2/2! + \lambda_4(T)\,t^4/4! + o(t^4) \qquad (18)$$

We then have, e.g.:

(a) The expected number of episodes in (t_0, t_f),

$$E\left[N_T(R, t_f - t_o)\right] = \lambda(t_f - t_o) \qquad (19)$$

where

$$\lambda = \left[\lambda_2(T)/2\pi\right]^{1/2}\,\phi(h) \qquad (20)$$

(b) The expected duration of an episode,

$$E\left[D_J\right] = \left[1 - \Phi(h)\right]/\lambda \qquad (21)$$

where

$$\Phi(h) = \int_{-\infty}^{h} \phi(x)\,dx \qquad (22a)$$

and

$$\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2) \qquad (22b)$$

Higher moments of these and related random variables have more complicated formulae. Simple asymptotic results are available if h is very large.

Integral functionals of $C_T(t)$ are particularly important for estimating cumulative dosage and dosage-response. The mean and variance of the so-called "Z_n-exceedance measures" are relatively easy to compute (Cramer and

Leadbetter (1967)). These are random variables defined by

$$Z_n(z,T) = \int_0^T \left[Z(t)-z\right]^n I_z\left[Z(t)\right] dt/T \tag{23}$$

where $I_z(Z)$ is the Heaviside step function defined in the previous section on Individual Dosage Histories. Thus

$$E\left[Z_n(z,T)\right] = \int_z^\infty (x-z)^n \phi(x)\, dx \tag{24}$$

and a more complicated formula for the variance. In particular, for:

(a) Lognormal distribution, $g(C) = \log_e C$

$$C_T(t) = \exp\left[\sigma Z(t) + \mu\right] \tag{25}$$

$$Z_n(z,T) = \sigma^{-n} \int_0^T \left[\log_e C_T(t) - \mu - z\sigma\right]^n \times I_{\exp(\mu + z\sigma)}\left[C_T(t)\right] dt/T \tag{26}$$

(b) Fourth-root Gaussian distribution, $g(C) = C^{1/4}$

$$C_T(t) = \left[\sigma Z(t) + \mu\right]^4 \tag{27}$$

$$Z_n(z,T) = \sigma^{-n} \int_0^T \left[C_T(t)^{1/4} - \mu - z\sigma\right]^n \times I_{(\mu+z\sigma)^4}\left[C_T(t)\right] dt/T \tag{28}$$

Note that if (as is often the case) physiological response is proportional to the logarithm of the stimulus above a certain theshold level $\mu + z\sigma$, then Equation 26 is most appropriate for health impact predictions.

These results are readily extended to curve crossings by non-stationary Gaussian processes. This generalization is needed to discuss the effects of pollution episode control strategies. An episode is declared on the basis of values of $C_T(t)$ and its time derivative $C_T'(t)$ over some interval, or equivalently, on the basis of the values of the jointly Gaussian process $Z(t)$, $Z'(t)$ during some interval. Once the episode is declared, the effect of the controls is to decrease the mean value with time. The correlation matrix of the controlled process is the matrix of partial correlations conditioned on the history of the process up to the time the control strategy is initiated. The results are rather complicated and details will be presented elsewhere.

The correlation structure of the process Z(t) is crucial in predicting episode frequency and duration, but is not well known. Merz, et al. (1972) find that in Los Angeles, there is a weak, yearly trend in weekly averages of hourly maxima for oxidants, CO, NO, and HC; superimposed on this are strong semi-annual (seasonal) and weekly regressions, and a possible bi-weekly regression for CO and NO. Ash, et al. (1972) find that the fourth-root transformation reduces the process $C_1(t)$ to a stationary Gaussian random process with approximately Markov dependence,

$$\rho_T(t) = \exp(-w|t|) \tag{29}$$

for T = 1 hour, CO and SO_2 in Camden and Bayonne in 1970-1971. Unfortunately, Equation 29 cannot apply for small times, since it implies that the process Z(t) is not differentiable. One possible solution is that Z(t) is doubly stochastic, and that w is itself a random variable. Larsen has observed (1971) that the standard deviation and the maxima of log $C_T(t)$ are very slowly decreasing power functions of T. This suggests (Marcus (1972)) an average correlation function for large times t of the form

$$\rho_0(t) = a|t|^{-b} + 0(|t|^{-b}) \qquad 0 < b < 1 \tag{30}$$

One distribution of w which effects this (by no means the only possibility) is a Gamma density with shape parameter b and scale parameter B,

$$p(w) = B^b\, w^{b-1} \exp(-Bw)/\Gamma(b) \tag{31}$$

whence

$$E\left[\rho_1(t)\right] = \int_0^\infty p(w) \exp(-w|t|)\, dw = B^b/(B+|t|)^b \tag{32}$$

which is similar to Equation 30.

I have worked out a numerical example (Marcus (1972)) for SO_2 in Washington, D. C., based on the lognormal model with an annual arithmetic average concentration of 0.05 ppm, logarithmic standard deviation $\sigma(0) = 0.760$, and

$$\rho_0(t) = 0.668\,|t|^{-0.10} \quad \text{for } |t| \text{ (hours)} > 1 \text{ hour} \tag{33}$$

The predictions are:

Standard	Averaging Time T, hr	Standard L_T, ppm SO_2	Avg. No. Yearly Exceedances	Avg. Episode Duration, hr
Primary	24	0.14	1.95	210
Secondary	3	0.50	0.20	7.84
Secondary	24	0.10	6.74	176

These predictions could, in principle, be compared with observations, e.g., Highway Research Board (1966). Unfortunately, my research has not been funded or supported and I do not personally have the time or computing resources needed to carry out the data analysis. The predicted values are typical and plausible.

The importance of the correlation structure of the process, and the formulation of health impact problems in terms of stochastic integrals, suggests that it may not be useful to study only the maxima of a series of independent random variables (e.g. Barlow (1971) and Singpurwalla (1972)). Maximum concentrations, and maxima of integral functionals, are of considerable interest, but it would be more informative to study them for non-stationary correlated random processes.

Applications to Human Populations: Some Problems

One of the first problems is that of defining an appropriate physiological response function for exposure to pollutant concentrations which vary with time. This depends significantly on the pollutant, the most sensitive organ, the time scale and method of elimination as well as other factors. These are discussed in a review by Saltzman (1970). See also Rossin and Roberts (1971).

The second problem is the personal trajectory estimation for different types of individuals. We first need to classify individuals by potential health effects (the preceding problem), and then to relate these to demographic characteristics of the individual—age, sex, income, state of health, occupation, etc. These demographic factors, and an inventory of land uses in the metropolitan region, largely determine the personal trajectory. The techniques for doing this are an essential part of travel demand forecasting (e.g., Hanafani (1972); Wooten and Pick (1967); Highway Research Board (1972)). A recent, very useful approach involves the estimation of personal trip patterns as a Markov chain (Sasaki (1972)).

Finally, meteorology and human activities interact in complicated ways not readily accessible to modeling. For example, on a hot, calm, overcast summer day (light winds and stable turbulence conditions being conducive to high pollutant concentrations), an unusually large number of people might absent themselves from downtown offices, reducing motor vehicles emissions downtown but possibly increasing them near roads leading to parks and beaches. Power plant emissions might also change as a result of the redistribution of air conditioning demands, etc. A severe winter storm would introduce a different set of interactions.

The interpolation and extrapolation of pollution in space and time from air quality data at fixed monitors is not an extremely difficult problem. The prediction of multivariate time series is well known (Cramer and Leadbetter (1967)). The extrapolation of spatial random fields can be conveniently done by the use of empirical eigenvectors Peterson (1970) (1972), thus also exposure along trajectories.

References

Highway Research Board, 1972: Transportation Demand and Analysis Techniques. 18 reports. *Highway Research Record No. 392.*

Ash, D., Bloomfield, P., and McNeil, D. R., 1972: On the statistical analysis of air pollution data. Statistics Dept., Princeton Univ., Princeton, N. J., Tech. Rept. 19, Ser. 2.

Barlow, R. E., 1971: Averaging time and maxima for air pollution concentrations. Univ. of Calif., Berkeley, Calif., Operations Research Center Rept. ORC-71-17.

Brook, R. R., 1972: The measurement of turbulence in a city environment. *J. Applied Meteorology. 11:* 443-450.

Cramer, H., and Leadbetter, M., 1967: *Stationary and Related Stochastic Processes.* John Wiley, New York, N. Y.

Hanafani, A. K., 1972: An aggregative model of trip making. *Transportation Research. 6:* 119-124.

Holzworth, G. C., 1967: Mixing depths, wind speeds and air pollution potential for selected locations in the United States. *J. Applied Meteorology. 6:* 1039-1044.

Larsen, R. I., 1971: *A Mathematical Model for Relating Air Quality Measurements to Air Quality Standards.* U. S. Env. Prot. Agency Publ. AP-89.

Luna, R. E., and Church, H. W., 1972: A comparison of turbulence intensity and stability ratio measurements to Pasquill stability classes. *J. Applied Meteorology. II:* 663-669.

Marcus, A. H., 1972: Air pollutant averaging times: Notes on a statistical model, and predicting the frequency and duration of air pollution emergencies: A statistical model. Johns Hopkins Univ., Baltimore, Md., Statistics Dept. Tech. Repts.

Merz, P. H., Painter, L. J., and Ryason, R. R., 1972: Aerometric data analysis-time series analysis and forecast and an atmospheric smog diagram. *Atmospheric Environment. 6:* 319-342.

Peterson, J. R., 1970: Distribution of sulfur dioxide over metropolitan St. Louis as described by empirical eigenvectors and its relation to meteorological parameters. *Atmospheric Environment. 4:* 501-518.

Peterson, J. T., 1972: Calculations of sulfur dioxide concentrations over metropolitan St. Louis. *Atmospheric Environment. 6:* 433-442.

Rossin, A. D., and Roberts, J. J., 1971: Episode control criteria and strategy for carbon monoxide. Paper 71-55, presented at 64th Annual Meeting of the Air Pollution Control Association.

Saltzman, B., 1970: Significance of sampling time in air monitoring. *J. Air Pollution Control Association. 20:* 660-665.

Sasaki, T., 1972: Estimation of person trip pattern through Markov chains. *Proc. Fifth International Symposium on Traffic flow Theory and Transportation, G. W. Newell* (ed.), American Elsevier, New York, N. Y.

Singpurwalla, N. D., 1972: Extreme values from a lognormal law with applications to air pollution problems. *Technometrics. 14:* 703-712.

Stern, A. C. (ed.), 1970: Proceedings of Symposium on Multiple-Source Urban Diffusion Models, U. S. Env. Prot. Agency Publ. AP-86.

U.S.E.P.A., 1971: Our Urban Environment and Our Most Endangered People. Report, U. S. Environmental Protection Agency.

U.S.P.H.S., 1966: *Continuous Air Monitoring Program, Washington, D. C., 1962-1963.* U. S. Public Health Service Publ. AP-23.

Wooten, H. J., and Pick, G. W., 1967: A model for trips generated by households. *J. Trans. Econ. Policy. 1:* 137-153.

DISCUSSION

Joseph Vasalli: I noticed that in a great many of the purely statistical papers that have been presented there's been a concern with the time variance of air pollution concentration. If you think of it in a slightly different fashion there is a spatial variation that is superimposed on the time variance such that if you are looking at a time-series average concentration over an area with time, you get a band instead of a line. I am wondering what is the effect of superimposing the spatial variance on the time variance?

Marcus: You mean a spatial variance in the individual movements or what? I don't quite follow that.

Visalli: If you attempt to sample over an area, instead of at one point . . .

Marcus: Yes. You are right. If you attempted to sample over an area instead of at a single point you would have a problem. That would be exactly analogous in a formal sense to the moving average problem that we've had in here. That is, if instead of an instantaneous spatial point concentration, you were somehow able to simultaneously accumulate measurements over a very large area, then you would have the two dimensional analog of the moving average sort of formulation we have here. You would have a moving average with respect not only to the time axis, but also to three dimensions. This sort of thing could be dealt with if you went over to, say, Gaussian random processes with multi-dimensional index sets, say, time and three space dimensions. In laboratory studies of turbulence, this kind of representation is necessary. Unfortunately it becomes very difficult to do anything in the spatially completely isotropic and homogeneous case, and this doesn't describe very many cities I know—Los Angeles, maybe. Even in Los Angeles there are places that are distinctive from other places.

Ralph Larsen: Your observation that if 1-hour concentrations of a pollutant are lognormal then the 24-hour observations by theory cannot be lognormal, but that the fit may be still good to lognormal, is confirmed in another field by R. L. Mitchell in the September, 1968, issue of the Journal of the Optical Society of America, in an article titled "Permanence of the lognormal distribution." His abstract states that the distribution of the sum of lognormal variates is shown for most cases of interest to be accurately represented by a lognormal distribution instead of a normal or Rayleigh distribution that might be expected from the Central Limit Theorem. Then he goes on to show in his analysis that he does tend to get summations which look lognormal, but they're not perfect, they're just quite close.

Marcus: I wasn't aware of that paper, I would be interested in seeing it. The lognormal distribution has a number of strange properties which haven't come out yet, which I think should be mentioned. It doesn't have a moment generating function that has a unique inverse, which is rather embarassing in some applications. As far as heavy tailed, I wasn't aware that the sums of lognormal variates don't coverge to normality very quickly. I suppose it shouldn't be too surprising. I'd like to believe in the Central Limit Theorem. But there is another family of heavy tailed distributions which hasn't been mentioned yet, the so-called stable distribution laws, which have a number of extremely awkward properties like having infinite variance and, in many cases, infinite mean values. On the other hand, besides being a very heavily skewed sort of distribution, they do have one good property and that is that sums, or more generally, moving averages, of stably distributed variates have a stable distribution law. This gives us a useful kind of reproductive property. The problem about dealing with distributions that don't have a theoretical finite variance I find rather horrifying and would prefer not to look into. The question of the underlying distribution structure is one I absolutely haven't discussed. I did it in the paper a little bit. I even tried to go into how, starting out from a fundamental diffusion model, you could try to derive a lognormal distribution either by assuming that some of the components of the concentration like reciprocal wind speed, and reciprocal product of the azimuthal times the elevation standard deviation times wind speed might be approximately lognormally distributed. But when you have a large number of point, line and area sources, as you do in an urban region, you have a combination of both multiplicative and additive factors which will give you a distribution which is not evidently lognormal or anything else for that matter, and I don't know how to handle that. I'm afraid you have a mixture of distributions with a heavy tail, and that's about all I can say.

Benarie: Being an engineer and not at all a mathematician I would have checked the theory by available monitors. In radiation contamination protection personal monitors that are movable with the person are very extensively used, which could be used as a check of your theory first of all. In industrial hygiene there are several types of portable particulate monitors and it should be checked on them.

Marcus: I thoroughly agree with that, and the analogy with the radiation dosimeters is perfect.

8. EVALUATING CONFORMITY WITH TWO-POINT AIR QUALITY STANDARDS, POLLUDEX*

HAROLD E. NEUSTADTER AND STEVEN M. SIDIK

Lewis Research Center
National Aeronautics and Space Administration
Cleveland, Ohio

and

JOHN C. BURR, JR.

Air Pollution Control Division
City of Cleveland, Ohio

Introduction

This report presents the results of various statistical analyses of data obtained by the Air Pollution Control Division (APCD) of Cleveland, Ohio. It contains a tabulation of averages, statistics relevant to lognormal distributions, and goodness-of-fit statistics. In addition, a pollution-level index is introduced which relates the measured pollution levels over a year to the existing air quality standards.

The air sampling program of APCD is currently in its sixth year. Twenty-four-hour samplings have been made of total suspended particulate (TSP) since January 1967, and of nitrogen dioxide (NO_2) and sulfur dioxide (SO_2) since January 1968. The sampling methods used are high-volume air sampling, Jacobs-Hochheiser, and West-Gaeke, respectively. The geographic deployment of sampling sites is shown in Figure 1. The meandering heavy line in the center of the city is the Cuyahoga River, about which is centered most of the region's heavy industry.

At present, there are 21 stations monitoring the air. Fifteen of these stations monitor all three pollutants, while the remaining six (stations O to T in Figure 1) measure TSP only. Seventeen of these sites have been in operation for more than 5 years. Stations B, D, K, and N have undergone relocation since their initial installation. However, because of the proximity of their present sites to their former sites, we have assumed that essentially the same environment has been measured throughout the 5-year period. Currently, the air is sampled every third day, although the sampling frequency has varied over the 5 years and has been as low as once-a-week. Some of these data have been presented elsewhere in a more preliminary manner (Neustadter, et al. (1972)). The data analysis reported herein was performed by the Environmental Research Office of the NASA Lewis Research Center (LeRC) as part of the preliminary phase of a joint APCD-LeRC program to study trace elements and compounds in airborne particulate matter.

---

*This paper has also been released as a LeRC publication, NASA TN D-6935 entitled "Statistical Summary and Trend Evaluation of Air Quality Data for Cleveland, Ohio, in 1967 to 1971: Total Suspended Particulate, Nitrogen Dioxide, and Sulfur Dioxide."

Cleveland Aerometric Data

Pertinent results are presented in Tables I, II, and III for TSP, NO_2, and SO_2, respectively. In each table, the first column gives an alphabetic designation of the monitoring site corresponding to the code shown in Figure 1. The second column lists the various parameters of interest for each of the pollutants. These parameters are (a) number of days observed (readings); (b) geometric (TSP) or arithmetic (SO_2 and NO_2) averages; (c) standard geometric deviation; (d) estimated value of the second largest pollution level for the year; and (e) an adjusted Kolmogorov-Smirnov goodness-of-fit statistic for lognormality, denoted as $(N)^{1/2}D$.

Air quality standards are set nationally by the Environmental Protection Agency (EPA) of the Federal Government (Anon. (1971)) and statewide by the Air Pollution Control Board of the Department of Health (DoH) of the State of Ohio (Ohio, (1972)). Whenever these two standards differ, we have chosen to work with the DoH (more stringent) standard, which is listed in the third column. In the remaining five columns are the various statistics for each of the years 1967 to 1971.

Number of Readings

For each pollutant, both EPA and DoH require a minimum of one sampling every sixth day, or an equivalent set of at least 61 random samples per year. Thus, we designate this standard as >60 in the tables. Even though early in the program some stations did not achieve 60 samples per year for each pollutant, we have included the analyses of these data sets in this report. At present, the nominal schedule of APCD calls for monitoring the environmental air every third day. In practice, this procedure generally allows sufficient margin for unanticipated disruptions (e.g., equipment failure) while still exceeding 60 readings per year.

Geometric and Arithmetic Averages

The geometric average is used in Table I, and the arithmetic average is used in Tables II and III. This corresponds to the particular averaging method stipulated by EPA and DoH standards. Calculations were performed whenever the number of readings exceeded 10. The values listed as standards are the DoH primary standards, which correspond to the EPA secondary standards.

Standard Geometric Deviation (SGD)

It has been noted that, irrespective of sampling duration or location, air sampling data are generally distributed lognormally (Larsen (1971)). When such is actually the case, the entire data set is sufficiently described by its geometric average and SGD. The higher the SGD, the greater the spread between the lower and higher values. As with the averages, SGD was calculated for data sets of more than 10 readings,

Second Largest Value

Both EPA and DoH standards for TSP and SO_2 specify that a certain level of pollution is ". . . not to be exceeded more than one time per year." This implies that for the 365 daily pollution levels per year (366 for leap years), there is no upper bound on the largest single level. However, the next largest value (i.e., the second most polluted day of the year) is required to be at or below the standard. Thus, Tables I, II, and III include estimates of the second highest pollution level for each year. As with the averages, the values listed here are the DoH primary standards, which correspond to EPA secondary standards. While NO_2 has only a standard for the annual average, we believe the estimated second largest level for a year is useful information and we have included it in Table II.

An approximation to the second largest pollution level estimate, for a year of n days, and a sample of N observations, is obtained by the following procedure. (The transformation to the logarithms of the data values is made because the expected values of normal order statistics are well developed in the literature, whereas we are not aware of any comparable development for lognormal distributions.) The logarithms $y_i = \ln(x_i)$ of the pollution levels x_i are computed. According to the assumption of lognormality, the y_i values follow a normal distribution. The sample mean $\bar{y}$ and sample standard deviation s_y of the set of logarithms are computed. From Harter (1961), the expected value of the second largest observation in a sample of 365 (366 in a leap year) independent values from a normal distribution is 2.63 (to three significant digits) standard deviations from the mean. This value, along with the average $\bar{y}$ and the standard deviation s_y of the set of logarithms, is used in the following equation to obtain the estimate of the second largest pollution level of the year:

$$y_{2nd} = \bar{y} + 2.63\, s_y \tag{1}$$

The values of x_{2nd} listed in Tables I, II, and III are obtained by exponentiation, as

$$x_{2nd} = \exp(y_{2nd}) \tag{2}$$

Because of the decreased precision which occurs when extrapolating to the tail of a distribution and because the sample mean and standard deviation are used, the minimum number of readings for this calculation was increased to 30 as opposed to 10 used for the averages. Implicit in using Equation 1 is the assumption of lognormality of the data, which leads us to the final entry in these tables.

Kolmogorov-Smirnov Statistic

The Kolmogorov-Smirnov statistic is a goodness-of-fit statistic which can be applied to any distribution (Noether (1967)). In testing for a lognormal distribution, it is easier for calculation purposes to take the logarithms of the values and test for goodness-of-fit to a normal distribution. This statistic was originally intended for use when the distribution which the data is suspected of following is completely specified. For the normal distribution, this is equivalent to knowing the mean μ and the standard deviation σ. In this case, the Kolmogorov-Smirnov statistic is denoted D and is calculated as

$$D = \max_{i=1,N} \left| \Phi\left[\frac{y_i - \mu}{\sigma}\right] - \left[\frac{i}{N}\right] \right| \tag{3}$$

where the function $\Phi(z)$ denotes the cumulative standard normal distribution function.

The statistic D measures the maximum deviation of the observed cumulative distribution function from the theoretical cumulative distribution function. Thus, D is always a value between 0 and 1. A value of 0 would indicate a perfect fit of the sampled data to a lognormal distribution, and larger values indicate an increasing deviation from lognormality.

When the mean and the standard deviation are unknown, it is common to use the estimates $\bar{y}$ and $s_y = [\Sigma_i(y_i - \bar{y})^2 /(N - 1)]^{1/2}$ in place of μ and σ. Lilliefors (1967) has studied the use of the Kolmogorov-Smirnov statistic in this situation. Table IV of this report presents the significance levels of $(N)^{1/2}D$ from Lilliefors (1967) for samples of $N > 30$. Thus, the statistics in Tables I, II, and III are presented as $(N)^{1/2}D$.

It should be recognized that the observed pollution levels are but a sample of levels from some distribution. Thus, even if the distribution of the complete set of pollution levels is indeed lognormal, some of the samples will lead to large values of $(N)^{1/2}D$. The interpretation of the tabulated significance levels α is that if the distribution is indeed lognormal, then about 100α percent of the samples tested will lead to a value of $(N)^{1/2}D$ which exceeds $((N)^{1/2}D)_\alpha$, whereas about $100(1 - \alpha)$ percent will lead to a value of $(N)^{1/2}D$ lower than $((N)^{1/2}D)_\alpha$. Because subsequent calculations in this report depend heavily on the assumption of lognormality, the value of $\alpha = 0.20$ was chosen. Choosing this large value for α has the drawback of rejecting the assumption of lognormality a substantial proportion of the times that the distribution is lognormal. However, it has the compensating advantage of being more discriminating against distributions which are not lognormal.

Lognormality

Lognormal Plots

As a graphical means of assessing the goodness-of-fit of the data to a lognormal distribution, we can enter the observed data on lognormal probability graphs. Figures 2 and 3 show two plots for TSP. The solid line indicates the plot of the cumulative sample distribution of all measurements over the 5-year period. The data points present the separate sample distributions for the 5 years (1967 to 1971). Any steady increase or decrease in the pollutant concentration would be discernible as a vertical sequence of the data points representing those years. In the two cases shown, there is no overall trend. Figure 2 is for station I in the industrial valley. The overprinting of the data points shows the TSP levels to be fairly uniform at a rather high average level for the 5-year period. Figure 3 represents station K, in a residential neighborhood, predominantly upwind from the industrial region.

A full set of lognormal curves for all 21 stations for the 3 pollutants is available on microfiche from the authors upon request.

Goodness of Fit

To indicate the decreasing likelihood of lognormality as $(N)^{1/2}D$ increases, all values calculated on the assumption of lognormality for which the goodness-of-fit statistic is outside the 20-percent confidence level (i.e., the data having $(N)^{1/2}D > 0.736$) are footnoted in the Tables. For a further indication of lognormality, as well as for a check on the consistency of our data, we examined the distribution of sets for which $(N)^{1/2}D > 0.736$.

Table V summarizes the results of the goodness-of-fit tests in which the a = 0.20 significance level was used. The first column lists the station identification. The remaining columns list for each of the pollutants the number of yearly tests which were performed, and the number of these tests which rejected the assumption of lognormality. For TSP, there are 85 tests, of which 20 were rejections. This is very close to the expected number of rejections and implies that the distribution of TSP may very safely be considered to be lognormal. For NO_2 and SO_2, however, there are more than twice as many rejections as would be expected, and hence their closeness to a lognormal distribution is somewhat suspect. On the basis of an examination of the lognormal plots of SO_2 and the fact that the SO_2 departure from lognormality, as indicated by $(N)^{1/2}D$, is not severe, we will proceed on the assumption that the lognormal is still a useful approximation to the distribution of SO_2.

Further examination of Table V shows that the lognormality of TSP, SO_2 and NO_2 is most questionable at stations E, F, and I. Benarie (1970) and Mitchell (1968) have each considered the additivity of lognormal distributions. Mitchell has shown that under certain conditions the sum of independent and identically distributed lognormal variates also follows a lognormal distribution. Benarie has considered a more general situation, where the lognormal variates have differing geometric means and standard geometric deviations. His conclusions are that when a large number (>10) of lognormal variates with slightly differing geometric means are superimposed, the resulting distribution is still well approximated by a lognormal distribution. However, when a small number (<10) of lognormal variates with differing means are superimposed, the resulting distribution generally is not a lognormal. Thus, it is possible to assume that pollution levels at stations E, F, and I are dominated by a small number of major sources, whereas the remaining stations reflect the influence of either a single large source or a superposition of many sources.

Air Quality

Among the goals of APCD are monitoring of the environmental air, determination of its quality, and initiation of action to improve the local air quality, where indicated. There are well established techniques for analyzing lognormal plots to extract information pertinent to determining compliance with air quality standards and/or the existence of long-term trends (Larsen (1971)). However, it is often desirable to have available some single number, or index, which presents as simply as possible a maximum of information. To this end we have developed an index, which we call Polludex, which gages the conformity of the measured environment to the established standards.

Polludex, An Air Pollution Index

Many indices have been proposed and a number are in use by various agencies (Babcock (1970)). Polludex is a variation of an index proposed by Pikul (1971). The rationale for constructing this modified index is as follows. The standards for TSP and SO_2 specify values for the annual mean which may not be exceeded and also values which may not be exceeded more than once per year. In relation to a lognormal plot of the underlying population, these standard values specify the coordinates of two points on a straight line. If the data obtained during a 1-year period conform to lognormality and conform to the required standards, the plot of the data will closely approximate a straight line falling entirely below (or on) the line segment joining the standard points.

For each of the three pollutants, define

$$r = \frac{\text{Sample average}}{\text{Standard for average}}$$

$$s = \frac{\text{Estimate of second largest level}}{\text{Standard not to be exceeded more than once yearly}}$$

Then Polludex, P (pollutant), is defined for TSP and SO_2 by

$$P(\mathrm{TSP}, SO_2) = 50 \times \left[\max(0, r-1) + \max(0, s-1)\right] \tag{4}$$

and for NO_2 by

$$P(NO_2) = 100 \times \left[\max(0, r-1)\right] \tag{5}$$

where max(a,b) means that the larger of the two values, a or b, is to be used. The geometric average is to be used in calculating r for TSP and the arithmetic average is to be used in calculating r for SO_2 and NO_2. For the estimate of the second largest level to be used for s, we used the approximate value listed in Table I for TSP and in Table III for SO_2.

With this definition, the same weight is given to the long-term (chronic) effects of pollution as is given to the severe short-term (episode) incident. The standards for these pollutants have presumably been set with regard to maximum acceptable levels for reasons of public health and/or welfare. Thus, we assume that normalization of the estimated mean and second highest values by the standards will, in a sense, put each P on an equal basis with respect to the potential harm caused by excesses. If the air quality is equal to or better than the standards, Polludex = 0. A value of Polludex = 100 can be understood to mean that the air is, in a sense, 100 percent polluted, in that a value of 100 is obtained when the average and the second highest values are each 100 percent higher than their respective permissible levels. Of course, Polludex = 100 would also result from a continuum of other combinations, as, for example, when the second highest value is three times its standard, provided the average was at or below its standard. Figure 4 graphically illustrates several of these possibilities. Figure 4(a) shows three possible examples which have P = 0. Figure 4(b) shows a line having P = 100, where both the mean and second largest standards are exceeded. Figure 4(c) shows a line where again P = 100, but the standard for the mean has been met. Finally, Figure 4(d) shows a line with P = 50, where the standard for the mean is not met but the other standard is.

Four-Year Trends

Polludex was evaluated for the APCD data and is listed for all three pollutants in Table VI. The State of Ohio standards were used in these calculations.

Where there are adequate data, the 1968 and 1971 values are also presented as bar graphs overprinted on the Cleveland map. The Polludex values for TSP, NO_2, and SO_2 are shown in Figures 5(a), (b), and (c), respectively. If there are two bars, the left bar represents 1968 and the right bar 1971. With the exception of site M of Figure 5(c), a single bar represents 1971. It is clear that, in general, TSP levels have increased to the west of the Cuyahoga River and decreased to the east. The most pronounced improvements are downwind of the valley (in Cleveland, the winds are predominantly out of the southwest) at sites A, I, and E. The levels of NO_2 show much less variation, except for the increased levels at sites H and C. With one exception, there has been a significant reduction in the levels of SO_2 throughout the city, with the most pronounced improvements occurring, as with TSP, at sites A, I, and E. Since space heating is fueled primarily by natural gas, this implies a reduction in SO_2 contamination by industrial and power-producing sources. At this time we do not have sufficient information to determine whether the improvements in the valley are due to the general decline in business activity in recent years, the abatement efforts by the industrial community, both of these reasons, or, possibly, neither of these reasons.

Concluding Remarks

Air quality data (total suspended particulate, nitrogen dioxide, and sulfur dioxide) for Cleveland, Ohio, for the period of 1968 to 1971 have been collated and subjected to statistical analysis. It is apparent that the data for total suspended particulate and, to a lesser degree, the data for sulfur dioxide and nitrogen dioxide are lognormally distributed. The air quality standards of the

State of Ohio are met only sporadically by sulfur dioxide in isolated residential neighborhoods. The available data indicate that definite improvement in air quality has taken place in the industrial region. Overall, there appears to be a net improvement in air quality, which would be a reflection primarily of the striking reduction in sulfur dioxide levels.

A pollution index has been introduced which directly displays information regarding the degree to which the environmental air conforms to the mandated standards for the environment. As such, it is a useful tool in air quality monitoring programs.

Table I. Total Suspended Particulate Data Summary for 1967 to 1971

Monitoring station, (see fig. 1)	Statistic	Standard	1967	1968	1969	1970	1971
A	Number of readings	>60	19	70	73	76	69
	Geometric average	60	190	242	199	188	183
	Standard geometric deviation		1.4	1.7	1.6	1.6	1.7
	Second highest reading	150		919	[a]711	[a]682	730
	Goodness-of-fit statistic, $(N)^{1/2}D$			0.53	0.84	0.81	0.73
B	Number of readings	>60	36	64	66	[b]72	63
	Geometric average	60	112	104	94	113	92
	Standard geometric deviation		1.5	1.6	1.4	1.6	1.6
	Second highest reading	150	351	349	226	370	319
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.76	0.72	0.63	0.48	0.53
C	Number of readings	>60	64	79	72	97	89
	Geometric average	60	124	121	107	124	121
	Standard geometric deviation		1.6	1.6	1.6	1.6	1.7
	Second highest reading	150	343	[a]429	346	420	502
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.55	0.76	0.50	0.39	0.65
D	Number of readings	>60	44	72	74	[b]62	[c]30
	Geometric average	60	134	126	123	154	163
	Standard geometric deviation		1.5	1.5	1.5	1.6	1.8
	Second highest reading	150	371	390	378	487	
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.37	0.42	0.50	0.40	
E	Number of readings	>60	61	75	75	93	80
	Geometric average	60	139	147	119	136	120
	Standard geometric deviation		1.4	1.5	1.4	1.5	1.5
	Second highest reading	150	352	[a]410	276	[a]395	[a]328
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.59	0.83	0.61	0.80	0.80
F	Number of readings	>60	64	75	75	82	74
	Geometric average	60	101	103	88	109	105
	Standard geometric deviation		1.5	1.6	1.6	1.5	1.5
	Second highest reading	150	303	357	297	307	304
	Goodness-of-fit statistic, $(N)^{1/2}D$		1.0	0.67	0.64	0.07	0.72
G	Number of readings	>60	8	75	73	103	83
	Geometric average	60		99	82	94	91
	Standard geometric deviation			1.6	1.6	1.7	1.6
	Second highest reading	150		317	[a]292	358	337
	Goodness-of-fit statistic, $(N)^{1/2}D$			0.56	0.79	0.59	0.57
H	Number of readings	>60		65	68	96	70
	Geometric average	60		83	84	94	89
	Standard geometric deviation			1.6	1.6	1.7	1.7
	Second highest reading	150		280	299	384	352
	Goodness-of-fit statistic, $(N)^{1/2}D$			0.53	0.59	0.48	0.68
I	Number of readings	>60	66	75	75	101	93
	Geometric average	60	210	232	223	225	196
	Standard geometric deviation		1.4	1.5	1.5	1.5	1.6
	Second highest reading	150	[a]543	694	[a]639	701	[a]658
	Goodness-of-fit statistic, $(N)^{1/2}D$		1.08	0.60	0.97	0.51	0.83

Table I (cont'd). Total Suspended Particulate Data Summary for 1967 to 1971

Monitoring station (see Fig. 1)	Statistic	Standard	1967	1968	1969	1970	1971
J	Number of readings	>60	63	76	74	103	90
	Geometric average	60	174	161	151	156	163
	Standard geometric deviation		1.5	1.8	1.7	1.6	1.7
	Second highest reading	150	474	[a]538	[a]613	[a]530	645
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.62	0.78	0.76	0.98	0.73
K	Number of readings	>60	74	75	[b]105	81	78
	Geometric average	60	53	58	59	49	92
	Standard geometric deviation		2.5	2.1	1.9	2.4	1.6
	Second highest reading	260	399	320	258	[a]359	312
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.55	0.57	0.64	0.83	0.52
L	Number of readings	>60			42	79	73
	Geometric average	60			157	116	212
	Standard geometric deviation				1.7	2.6	1.6
	Second highest reading	260			569	[a]1013	637
	Goodness-of-fit statistic, $(N)^{1/2}D$				0.62	0.98	0.64
M	Number of readings	>60	53	73	98	58	72
	Geometric average	60	50	55	58	41	82
	Standard geometric deviation		1.9	1.9	2.3	2.8	1.6
	Second highest reading	260	220	235	309	[a]372	284
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.72	0.67	0.67	0.74	0.59
N	Number of readings	>60			35	81	86
	Geometric average	60			68	72	138
	Standard geometric deviation				2.6	2.9	2.0
	Second highest reading	260			[a]548	[a]755	905
	Goodness-of-fit statistic, $(N)^{1/2}D$				0.78	0.90	0.71
O	Number of readings	>60	69	75	72	90	76
	Geometric average	60	92	86	79	89	90
	Standard geometric deviation		1.5	1.6	1.6	1.7	1.8
	Second highest reading	150	265	298	[a]270	333	422
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.62	0.39	0.83	0.71	0.55
P	Number of readings	>60	62	74	72	93	74
	Geometric average	60	135	139	127	137	146
	Standard geometric deviation		1.4	1.5	1.6	1.5	1.4
	Second highest reading	150	343	390	407	412	371
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.71	0.40	0.64	0.55	0.60
Q	Number of readings	>60	63	69	70	88	79
	Geometric average	60	105	95	96	106	101
	Standard geometric deviation		1.5	1.5	1.4	1.8	1.4
	Second highest reading	150	310	277	241	[a]405	256
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.82	0.42	0.67	0.97	0.65
R	Number of readings	>60	57	72	65	90	66
	Geometric average	60	81	80	81	89	89
	Standard geometric deviation		1.6	1.7	1.6	1.6	1.7
	Second highest reading	150	265	304	265	309	384
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.44	0.69	0.52	0.49	0.60
S	Number of readings	>60					51
	Geometric average	60					92
	Standard geometric deviation						1.5
	Second highest reading	150					290
	Goodness-of-fit statistic, $(N)^{1/2}D$						0.71

Table I (cont'd). Total Suspended Particulate Data Summary for 1967 to 1971

Monitoring station (see Fig. 1)	Statistic	Standard	1967	1968	1969	1970	1971
T	Number of readings	>60					41
	Geometric average	60					170
	Standard geometric deviation						2.0
	Second highest reading	150					1014
	Goodness-of-fit statistic, $(N)^{1/2}D$						0.48
U	Number of readings	>60					[d]34
	Geometric average	60					114
	Standard geometric deviation						2.3
	Second highest reading	150					137
	Goodness-of-fit statistic, $(N)^{1/2}D$						0.55

[a] The calculation used to obtain this estimate assumed lognormality despite $(N)^{1/2}D \geq 0.736$.

[b] Sampling site was relocated within same general neighborhood in midyear. It is assumed that for sampling purposes the environmental air was the same at both locations.

[c] Temporarily discontinued because of construction at sampling site.

[d] Sampling was initiated in the latter part of the year.

Table II. Nitrogen Dioxide Data Summary for 1968 to 1971

Monitoring station (see Fig. 1)	Statistic	Standard	1968	1969	1970	1971
A	Number of readings	>60	71	73	84	86
	Geometric average	100	211	220	214	202
	Standard geometric deviation		1.4	1.4	1.4	1.5
	Second highest reading		517	470	464	538
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.60	0.57	0.61	0.59
B	Number of readings	>60			9	81
	Geometric average	100				190
	Standard geometric deviation					1.5
	Second highest reading					[a]539
	Goodness-of-fit statistic, $(N)^{1/2}D$					0.77
C	Number of readings	>60	76	75	115	96
	Geometric average	100	177	248	234	255
	Standard geometric deviation		1.5	1.3	1.4	1.6
	Second highest reading		[a]495	[a]454	[a]576	835
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.87	0.88	0.88	0.64
D	Number of readings	>60	55	70	[b]83	[c]47
	Geometric average	100	207	219	217	205
	Standard geometric deviation		1.4	1.3	1.5	1.6
	Second highest reading		497	424	576	686
	Goodness-of-fit statistic, $(N)^{1/2}D$		1.65	0.70	1.03	0.62
E	Number of readings	>60	69	74	108	96
	Geometric average	100	203	237	217	205
	Standard geometric deviation		1.4	1.3	1.4	1.6
	Second highest reading		497	[a]437	[a]504	[a]686
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.70	0.90	1.39	1.69
F	Number of readings	>60	47	74	96	86
	Geometric average	100	212	197	215	203
	Standard geometric deviation		1.4	1.3	1.3	1.5
	Second highest reading		[a]511	[a]370	444	[a]518
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.78	0.76	0.70	0.93

Table II (cont'd). Nitrogen Dioxide Data Summary for 1968 to 1971

Monitoring station (see Fig. 1)	Statistic	Standard	1968	1969	1970	1971
G	Number of readings	>60	72	72	104	89
	Geometric average	100	201	221	224	203
	Standard geometric deviation		1.5	1.3	1.3	1.5
	Second highest reading		371	[a]432	453	516
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.56	0.91	0.43	0.65
H	Number of readings	>60	66	71	114	78
	Geometric average	100	166	225	213	202
	Standard geometric deviation		1.5	1.3	1.4	1.6
	Second highest reading		[a]471	[a]443	464	[a]633
	Goodness-of-fit statistic, $(N)^{1/2}D$		1.03	0.75	0.70	1.1
I	Number of readings	>60	67	76	111	88
	Geometric average	100	247	253	238	217
	Standard geometric deviation		1.4	1.3	1.3	1.5
	Second highest reading		535	495	[a]495	[a]615
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.45	0.71	1.1	0.93
J	Number of readings	>60		52	113	93
	Geometric average	100		225	255	240
	Standard geometric deviation			1.4	1.4	1.5
	Second highest reading			488	[a]548	600
	Goodness-of-fit statistic, $(N)^{1/2}D$			0.65	0.82	0.58
K	Number of readings	>60	74	74	[b]104	88
	Geometric average	100	162	192	209	183
	Standard geometric deviation		1.5	1.4	1.4	1.6
	Second highest reading		433	417	[a]486	565
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.53	0.67	0.76	0.67
L	Number of readings	>60			41	80
	Geometric average				220	219
	Standard geometric deviation				1.4	1.5
	Second highest reading				513	572
	Goodness-of-fit statistic, $(N)^{1/2}D$				0.68	0.71
M	Number of readings	>60	55	74	96	73
	Geometric average		157	168	176	159
	Standard geometric deviation		1.4	1.3	1.3	1.6
	Second highest reading		[a]342	335	341	507
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.80	0.60	0.65	0.54
N	Number of readings	>60			39	88
	Geometric average				208	223
	Standard geometric deviation				1.6	1.6
	Second highest reading				647	[a]712
	Goodness-of-fit statistic, $(N)^{1/2}D$				0.65	0.95
U	Number of readings	>60				[d]36
	Geometric average	100				230
	Standard geometric deviation					1.9
	Second highest reading					[a]1030
	Goodness-of-fit statistic, $(N)^{1/2}D$					1.34

[a] The calculation used to obtain this estimate assumed lognormality despite $(N)^{1/2}D \geqslant 0.736$.

[b] Sampling site was relocated within same general neighborhood in midyear. It is assumed that for sampling purposes the environmental air was the same at both locations.

[c] Temporarily discontinued because of construction at sampling site.

[d] Sampling was initiated in the latter part of the year.

Table III. Sulfur Dioxide Data Summary for 1968 to 1971

Monitoring station (see Fig. 1)	Statistic	Standard	1968	1969	1970	1971
A	Number of readings	>60	71	74	82	88
	Arithmetic average	60	137	135	116	84
	Standard geometric deviation		2.4	2.0	1.9	2.2
	Second highest reading	260	[a]972	[a]674	[a]518	523
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.75	0.96	0.88	0.66
B	Number of readings	>60			9	86
	Arithmetic average	60				50
	Standard geometric deviation					2.1
	Second highest reading	260				284
	Goodness-of-fit statistic, $(N)^{1/2}D$					0.70
C	Number of readings	>60	72	76	105	93
	Arithmetic average	60	95	85	74	67
	Standard geometric deviation		2.4	2.3	2.3	2.4
	Second highest reading	260	644	546	476	485
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.61	0.48	0.54	0.73
D	Number of readings	>60	53	72	[b]79	[c]45
	Arithmetic average	60	106	103	109	89
	Standard geometric deviation		1.8	1.7	2.0	2.0
	Second highest reading	260	413	278	[a]538	[a]469
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.52	0.47	0.91	0.76
E	Number of readings	>60	71	75	107	94
	Arithmetic average	60	112	107	96	65
	Standard geometric deviation		1.9	1.6	1.8	2.1
	Second highest reading	260	476	314	[a]397	375
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.68	0.42	0.88	0.71
F	Number of readings	>60	47	75	97	86
	Arithmetic average	60	84	76	90	59
	Standard geometric deviation		1.9	2.1	1.8	2.3
	Second highest reading	260	[a]364	[a]409	373	[a]401
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.80	1.04	0.68	0.83
G	Number of readings	>60	69	71	105	86
	Arithmetic average	60	77	58	63	50
	Standard geometric deviation		2.1	2.0	1.9	2.4
	Second highest reading	260	414	294	295	[a]363
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.57	0.70	0.70	0.75
H	Number of readings	>60	62	71	113	72
	Arithmetic average	60	64	63	66	48
	Standard geometric deviation		2.3	2.3	2.2	2.4
	Second highest reading	260	[a]416	390	408	336
	Goodness-of-fit statistic, $(N)^{1/2}D$		0.85	0.69	0.47	0.72
I	Number of readings	>60	64	77	108	83
	Arithmetic average	60	129	110	101	67
	Standard geometric deviation		1.8	1.8	1.9	2.1
	Second highest reading	260	[a]522	467	[a]449	[a]358
	Goodness-of-fit statistic, $(N)^{1/2}D$		1.04	0.64	0.87	0.90
J	Number of readings	>60		32	113	93
	Arithmetic average	60		113	124	79
	Standard geometric deviation			1.9	1.8	2.0
	Second highest reading	260		543	504	[a]410
	Goodness-of-fit statistic, $(N)^{1/2}D$			0.53	0.70	1.23

Table III (cont'd). Sulfur Dioxide Data Summary for 1968 to 1971

Monitoring station (see Fig. 1)	Statistic	Standard	1968	1969	1970	1971
K	Total suspended particulate	[a]55	[a]59	43	[b]59	81
	Nitrogen dioxide		62	92	[b]109	83
	Sulfur dioxide		27	11	[b]0	[a]19
L	Total suspended particulate				222	280
	Nitrogen dioxide				120	119
	Sulfur dioxide				141	[a]192
M	Total suspended particulate	61	62	37	70	63
	Nitrogen dioxide		57	68	76	59
	Sulfur dioxide		0	0	9	[a]22
N	Total suspended particulate	206	293	268	[b]436	317
	Nitrogen dioxide				108	[a]127
	Sulfur dioxide				[a]62	[a]105
O	Total suspended particulate	65	71	[a]56	85	116
P	Total suspended particulate	127	146	142	151	145
Q	Total suspended particulate	91	71	60	[a]153	69
R	Total suspended particulate	56	68	62	77	102
S	Total suspended particulate					73
T	Total suspended particulate					380
U	Nitrogen dioxide					[d]129
	Sulfur dioxide					[d]138

[a] The calculation used to obtain this estimate assumed lognormality despite $(N)^{1/2}D \geqslant 0.736$.

[b] Sampling site was relocated within same general neighborhood in midyear. It is assumed that for sampling purposes the environmental air was the same at both locations.

[c] Temporarily discontinued because of construction at sampling site.

[d] Sampling was initiated in the latter part of the year.

Table IV. - Significance Levels for the
Kolmogorov-Smirnov Goodness-of-Fit Statistic
[From Lilliefors (1967)]

Significance level, α	0.20	0.15	0.10	0.05	0.01
Statistic, $(N)^{1/2}D_\alpha$	0.736	0.768	0.805	0.886	1.031

Table V. - Summary of Results of Goodness-of-Fit Tests

Monitoring station (see fig. 1)	Total suspended particulate		Nitrogen dioxide		Sulfur dioxide	
	Number of tests	Rejected	Number of tests	Rejected	Number of tests	Rejected
A	4	2	4	0	4	3
B	5	0	1	1	1	0
C	5	1	4	3	4	0
D	4	0	4	2	4	2
E	5	3	4	3	4	1
F	5	2	4	3	4	3
G	4	1	4	1	4	1
H	4	0	4	3	4	1
I	5	3	4	2	4	3
J	5	3	3	1	3	1
K	5	2	4	1	4	1
L	2	0	2	0	2	1
M	5	0	4	1	4	1
N	5	1	2	1	2	2
O	5	1				
P	5	0				
Q	5	1				
R	5	0				
S	1	0				
T	1	0				
U			1	1	1	0
Total	85	20	49	23	49	20
Percentage of tests rejected		24		47		41
Expected number of rejections		17		9.8		9.8

Table VI. Polludex Values for 1967 to 1971.

Monitoring station (see fig. 1)	Pollutant	1967	1968	1969	1970	1971
A	Total suspended particulate		408	[a]303	[a]284	296
	Nitrogen dioxide		111	120	114	102
	Sulfur dioxide		[a]201	[a]142	[a]97	70
B	Total suspended particulate	111	103	54	[b]117	82
	Nitrogen dioxide					90
	Sulfur dioxide					5
C	Total suspended particulate	117	[a]144	105	144	167
	Nitrogen dioxide		77	148	134	155
	Sulfur dioxide		103	75	55	49
D	Total suspended particulate	135	135	129	[b]191	(c)
	Nitrogen dioxide		107	119	[b]117	[c]99
	Sulfur dioxide		68	58	[a,b]94	[a,c]64
E	Total suspended particulate	133	[a]159	91	[a]145	[a]109
	Nitrogen dioxide		103	137	117	105
	Sulfur dioxide		85	50	[a]56	26
F	Total suspended particulate	[a]85	104	72	[a]93	89
	Nitrogen dioxide		112	97	115	103
	Sulfur dioxide		[a]40	[a]42	47	27
G	Total suspended particulate		89	[a]66	98	89
	Nitrogen dioxide		101	121	124	103
	Sulfur dioxide		44	7	10	[a]20
H	Total suspended particulate		62	70	106	91
	Nitrogen dioxide		66	125	113	102
	Sulfur dioxide		[a]34	27	34	15
I	Total suspended particulate	[a]255	324	[a]299	321	[a]283
	Nitrogen dioxide		147	153	138	117
	Sulfur dioxide		[a]108	82	[a]70	25
J	Total suspended particulate	203	[a]213	[a]230	[a]207	251
	Nitrogen dioxide			125	155	140
	Sulfur dioxide			99	100	[a]45

[a] The calculation used to obtain this estimate assumed lognormality despite $(N)^{1/2}D \geq 0.736$.

[b] Sampling site was relocated within same general neighborhood in midyear. It is assumed that for sampling purposes the environmental air was the same at both locations.

[c] Temporarily discontinued because of construction at sampling site.

[d] Sampling was initiated in the latter part of the year.

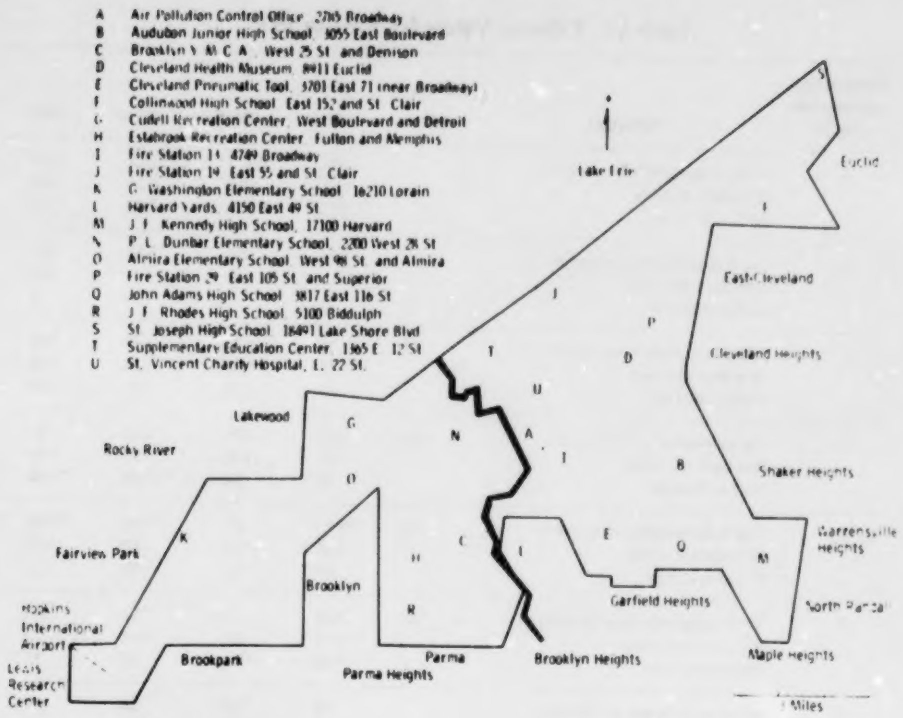

A Air Pollution Control Office 2765 Broadway
B Audubon Junior High School 3055 East Boulevard
C Brooklyn Y M C A West 25 St and Denison
D Cleveland Health Museum 8911 Euclid
E Cleveland Pneumatic Tool 3701 East 71 (near Broadway)
F Collinwood High School East 152 and St Clair
G Cudell Recreation Center West Boulevard and Detroit
H Estabrook Recreation Center Fulton and Memphis
I Fire Station 13 4749 Broadway
J Fire Station 19 East 55 and St Clair
K G Washington Elementary School 16210 Lorain
L Harvard Yards 4150 East 49 St
M J F Kennedy High School 17100 Harvard
N P L Dunbar Elementary School 2200 West 28 St
O Almira Elementary School West 98 St and Almira
P Fire Station 29 East 105 St and Superior
Q John Adams High School 3817 East 116 St
R J F Rhodes High School 5100 Biddulph
S St Joseph High School 18491 Lake Shore Blvd
T Supplementary Education Center 1365 E 12 St
U St. Vincent Charity Hospital, E. 22 St.
Lake Erie
Euclid
East Cleveland
Cleveland Heights
Shaker Heights
Warrensville Heights
North Randall
Maple Heights
Garfield Heights
Brooklyn Heights
Parma
Parma Heights
Brookpark
Brooklyn
Lakewood
Rocky River
Fairview Park
Hopkins International Airport
Lewis Research Center
Miles

Figure 8-1. Air pollution monitoring sites for Cleveland, Ohio. The heavy line down the center is the Cuyahoga River. The municipal boundaries have been straightened somewhat but are accurate in their essential features.

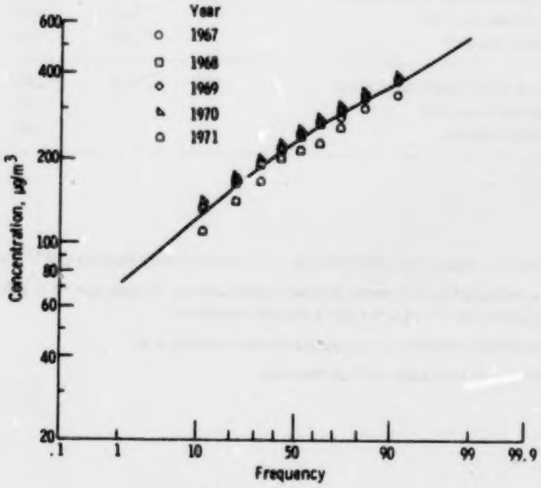


Figure 8-2. Lognormal plot of distribution by weight of total suspended particulate (24-hr. sampling) for monitoring station I (see Fig. 1) downwind of the industrial region.

Reproduced from best available copy.

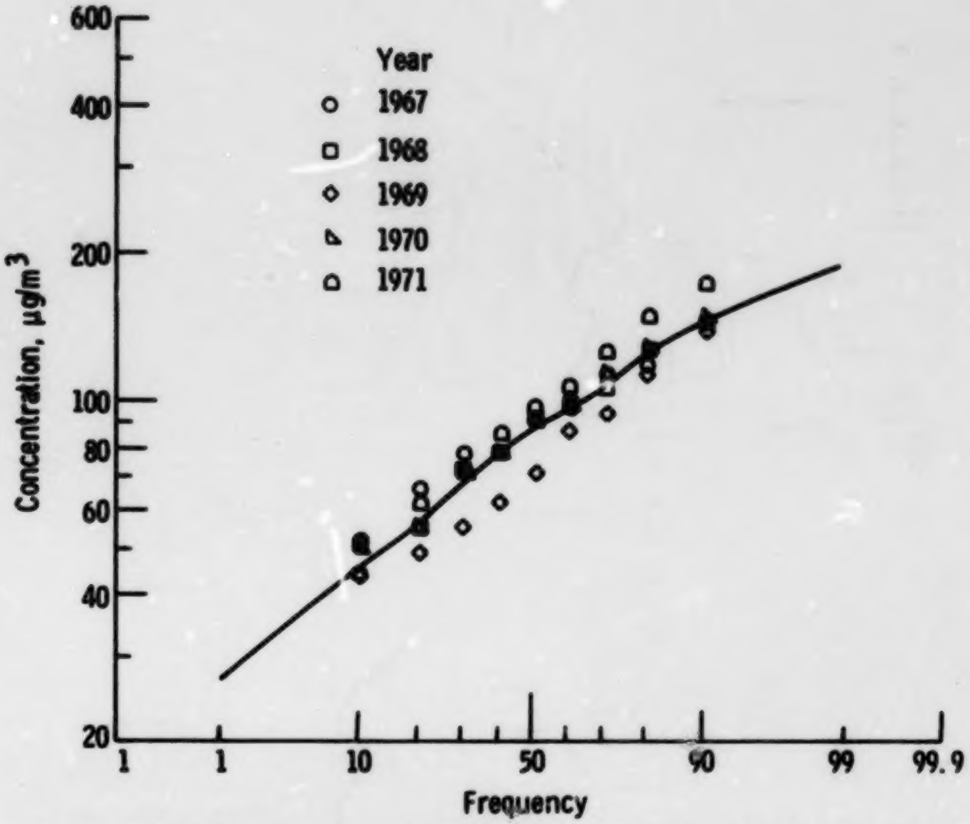
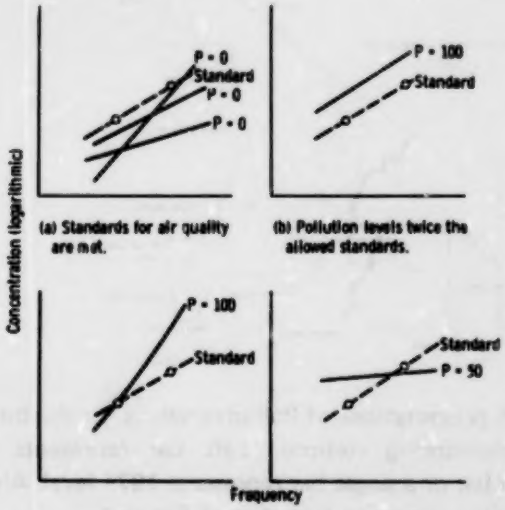


Figure 8-3. Lognormal plot of distribution by weight of total suspended particulate (24-hr. sampling) for monitoring station K (see Fig. 1) upwind of the industrial region.



(a) Standards for air quality are met.

(b) Pollution levels twice the allowed standards.

Figure 8-4. Examples of Polludex levels.

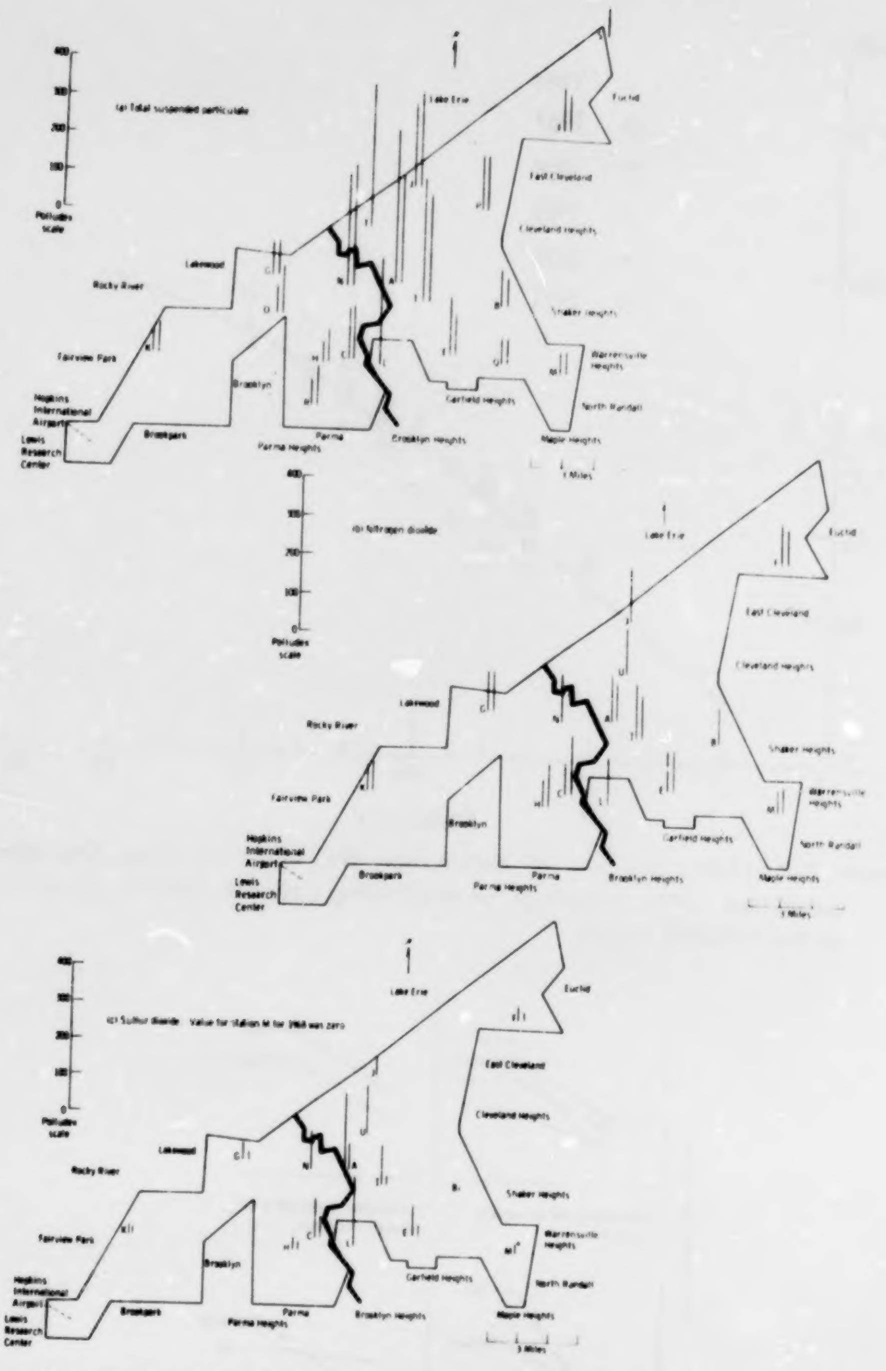

(a) Total suspended particulate
(b) Nitrogen dioxide
(c) Sulfur dioxide. Value for station M for 1968 was zero
Polludex scale
0
100
200
300
400
Lake Erie
Euclid
East Cleveland
Cleveland Heights
Shaker Heights
Warrensville Heights
North Randall
Maple Heights
Garfield Heights
Brooklyn Heights
Parma
Parma Heights
Brooklyn
Brookpark
Lakewood
Rocky River
Fairview Park
Hopkins International Airport
Lewis Research Center
3 Miles

Figure 8-5. Bar graph presentations of Polludex values for the three pollutants at the various monitoring stations. Left bar represents 1968 level of pollution; right bar or a single bar represents 1971 level. Alphabetic coding of monitoring sites corresponds to that of Figure 1.

Reproduced from best available copy.

References

Anon., 1971: *Federal Register. 36*: 8186.

Babcock, L. R., 1970: A Combined Pollution Index for Measurement of Total Air Pollution, *J. Air Pollution Control Association. 10:* 653.

Benarie, M., 1971: Sur la Validitie de la Distribution Logarithmical-normalie des Concentrations de Polluant. *Proc. Second Intern. Clean Air Congress,* pp. 68-70, eds., H.M. Englund and W. T. Beery, Academic Press, New York.

Harter, H. L., 1961: Expected Values of Normal Order Statistics. *Biometrika. 48:* 151-165.

Larsen, R. I., 1971: A Mathematical Model for Relating Air Quality Measurements to Air Quality Standards. Environmental Protection Agency, Office of Air Programs, U. S. Rep. AP-89.

Lilliefors, H. W., 1967: On the Kolmogorov-Smirnov Test for Normality With Mean and Variance Unknown. *J. Am. Stat. Assoc. 62:* 399-402.

Mitchell, R. I., 1968: Permanence of the Log-Normal Distribution. *J. Opt. Soc. Am. 58:* 1267-1272.

Neustadter, H. E.; King, R. B.; Fordyce, J. S.; and Burr, J. C., Jr., 1972: Air Quality Aerometric Data for the City of Cleveland from 1967 to 1970 for Sulfur Dioxide, Suspended Particulates, and Nitrogen Dioxide. NASA TM X-2496.

Noether, G. E., 1967: Elements of Nonparametric Statistics. John Wiley & Sons, Inc. New York, N. Y.

Ohio, 1972: State of Ohio, Department of Health, Air Pollution Unit. Regulations AP-3-02, AP-7-01.

Pikul, R., 1972: Development of Environmental Indices. Mitre Corp., Rep. M71-47.

DISCUSSION

Singpurwalla: When you tried these Kolmogorov-Smirnov goodness of fit tests could you care to tell which tables you used for the levels of significance?

Neustadter: I believe they are referenced in the report. I don't have it in my head. The statisticians did it, but I believe the tables are in the report and the reference is there.

Singpurwalla: The reason for questioning this is because if you estimate the parameters from the data and go ahead and use Kolmogorov-Smirnov tables which are generally available, then you are apt to make some kind of an error. But there are modified tables.

Neustadter: We are aware of that. We used the modified table.

Singpurwalla: O.K., and I was wondering if that might be any reason why it might change the answer.

Neustadter: No. We are aware of the problem of using estimated parameters and we did use modified tables. [See Lilliefors (1967).]

Rustagi: This morning I have heard quite a bit of glorification of lognormal distribution. I want to add one more reference to that. In 1964, *Archives of Environmental Health* I have the paper, titled "Stochastic behavior of trace substances," in which many of these substances have been studied including air pollutants. The amazing thing was that many substances were in liquids. For example, amino acids in urine also followed lognormal distribution. The second point I want to ask is for Mr. Marcus, who used the concept of a very interesting cumulative dose. I think most of the trace metals such as lead, about which I am familiar with, in the human body or biological systems are excreted in also certain random fashion. Could the deposition of a substance like lead or other gases be put into the model? A simple model in that connection was also mentioned by me in *Archives of Environmental Health* giving a model of body burden where intake and output were used in the model, however, not as the formal stochastic processes, rather as probability distributions without any assumptions for parametric form such as lognormal. I would like to mention the physiological experiments connected with air pollutants. There are very few studies but I think the audience should be aware of two famous studies—one is on human subjects over the past thirty years on lead by Dr. R. A. Kehoe and I think it is given in a series of lectures by Dr. Kehoe, "Metabolism of lead in man in health and disease," where he studied whatever metabolism could be studied in man. In animals Professor Herman Cember of Northwestern has studied the metabolism of mercury in rats over a period of time and I think these two studies should be noted. I'm not aware of others.

9. AN INVESTIGATION OF THE FREQUENCY DISTRIBUTIONS OF SURFACE AIR-POLLUTANT CONCENTRATIONS

J. B. KNOX

R. I. POLLACK

Lawrence Livermore Laboratory
University of California, Livermore, California

In several papers, Larsen and co-workers (1965; 1967; 1969) have discussed the frequency distributions of various pollutant concentrations calculated from data taken at CAMP (Continuous Air Monitoring Program) sites for 3 years in various cities. The data consist of instantaneous measurements taken at 5-minute intervals. When used in this fashion, or averaged over any period of time, the resulting frequency distribution is in all cases approximately lognormal. It was also noted that median concentration is proportional to averaging time to an exponent.

This result allowed these investigators to relate the geometric mean, standard geometric deviation and averaging time to the probable number of times during a year that a given level of pollution would be exceeded. This type of data presentation is useful because ambient air quality standards are set in the form of a maximum allowable average over a given period of time e.g., 0.03 ppm for 8-hr-duration samples would be allowed once a year for oxidant.

The CAMP data indicates that reactive pollutants may have a larger SGD (standard geometrical deviation) because of the additional variability introduced by the nature of chemical or photochemical reactions. It was noted by Larsen and verified by Knox and Lange (1972) that continuous point sources give concentration distributions with larger SGD's than area sources. This is attributed to the dependence of the pollutant concentration on the lateral and vertical standard deviation of the plume.

Barlow (1971) suggested that the lognormal distribution may not be appropriate because the averaging process implies that the sum of lognormally distributed random variables is itself lognormal. This is contrary to statistical theory. He suggests that a Weibull distribution would be more appropriate. This suggestion is supported by Milokaj (1972) who has argued the validity of the Weibull distribution for a variety of situations involving pollutant concentrations. He emphasizes the importance of the threshold parameter (γ).

The probability of occurrence of a value of the random variable smaller than γ is essentially zero. The lognormal distribution also has a formulation including a threshold parameter, however it is usually ignored due to the fact that low concentrations are ordinarily beneath the sensitivity of the measuring instruments. Also, the most interesting cases are the higher concentration levels where the lognormal fits well, and the significance of a threshold parameter is minimal. There is some indication, justifying further investigation, that the lognormal fits well at both high and low concentration levels, but with slightly different parameters. This suggests that two adjacent lognormal distributions may be present, perhaps caused by different types of meteorology.

The motivation for using the Weibull distribution is largely empirical. This distribution, with density function

$$f(x) = kx^{m} e^{-kx^{m+1}/(m+1)}$$

is widely used for a failure law for systems composed of a number of components where failure is due to the most severe flaw among many in the system. Barlow notes that histograms of air pollution concentrations are similar to life test data plots, which have been fit well with the Weibull distribution. The fundamental reason for pollutant concentrations to follow this distribution is not clear. Indeed, in his original paper, Weibull (1951) notes that there is no theoretical basis for this distribution. Figures 1 and 2 show, for illustrative purposes, the Weibull and lognormal probability plots for hourly averages of CO in San Francisco. It is interesting that several other distributions have been used with reasonable success, e.g., the gamma and beta distributions, but no theoretical basis has been proposed for these either.

Singpurwalla (1972) has interpreted the lognormal distribution using extreme value theory. He developed the limiting distribution of the maximum term of a random series from the lognormal distribution. This result can be used to compare pollutant concentration distributions to air quality standards. However, one of the assumptions required is that the series of observations be independent. He suggests that this is the case for averaging times of 8 hrs or more. The validity of this assumption is questioned later in this paper.

The Lognormality of Meteorological Variables

Other studies have shown that a variety of meteorological variables are also lognormally distributed.

The distribution of particle sizes in atmospheric aerosols has been found to be lognormal at a reasonable level of significance by several investigators (Blifford and Gillette (1971); Friedlander (1960)). The reason for this lies in the physical mechanisms by which particles are introduced and removed from the aerosol. Very simply, coagulation of particles smaller than the minimum investigated, forms particles which are measured. These form still larger particles until their increased weight causes them to fall out. This growth process dictates that the size of a particle is a function of its previous size multiplied by the rate of coagulation—a multiplicative process. This can be shown to yield a lognormal distribution. The rate of coagulation is a function of a variety of other variables.

Another important variable which has been found to be lognormally distributed is the rate of dissipation of energy in turbulence (ϵ). This distribution was postulated by Kolmogorov (1941) based on the assumption of a cascade process by which energy is transferred from a large-scale turbulent motion to progressively smaller-scale motions. The transfer stages are assumed similar and independent. Thus, the amount of energy dissipated at any stage is a function of the amount dissipated at the previous stage. The process may be viewed as

$$\epsilon_i = (\epsilon_{i-1}) Y_i \tag{1}$$

where ϵ_i is the energy dissipated at stage i and Y_i is a characteristic of the transfer stage causing the change in ϵ_{i-1}.

Due to the reproductive properties of the lognormal distribution, the distribution of ϵ implies the lognormality of several related variables including the dissipation of temperature by thermal eddy conduction, the squared-space differences of temperature and velocity, which imply that the differences themselves are lognormally distributed on either side of the origin, and the horizontal eddy diffusivity. These imply that the diffusive transfer between adjacent volumes of air is lognormally distributed. Furthermore, the lognormality of wind speeds, which has been demonstrated empirically, implies that the advective transfer is also lognormally distributed. These distributions have all been verified experimentally (Knox and Lange (1972); Gibson, et al. (1970) and (1970a)).

The Lognormal Process

The fundamental question has yet to be answered: Why are all these variables lognormally distributed? What underlying physical phenomena cause the lognormality of these variables? The answer to these questions requires a basic understanding of the theory behind the generation of the lognormal distribution.

Consider a stochastic process of the form

$$X_i = X_{i-1} + X_{i-1} Y_i \tag{2}$$

where Y_i is an independent stochastic variable, arbitrarily distributed.

If we solve Equation 2 for Y_i.

$$\frac{X_i - X_{i-1}}{X_{i-1}} = Y_i$$

and sum both sides,

$$\sum_{0}^{N} \frac{X - X_{i-1}}{X_{i-1}} = \sum_{0}^{N} Y_i$$

We can approximate* the left side by

$$\int_{n=0}^{N} \frac{dx}{X} = \sum_{0}^{N} Y_i$$

$$\ln X_N / X_0 = \sum_{0}^{N} Y_i$$

by the Central Limit Theorem $\sum_{0}^{N} Y_i$ is normally distributed, hence X_n is lognormally distributed.

This is known as the law of proportional effect; the percentage change in a variable is equal to a constant plus an error. If the absolute change had been equal to this same constant + error term, the normal distribution would have resulted. Hence, the lognormal distribution is the result of a multiplicative process whereas the normal distribution results from an additive process.

The basic properties of the lognormal are all multiplicative analogies to the normal distribution. This includes the reproductive properties. In particular the product of two lognormal distributions is lognormal, the sum, however, is not. Aitchison and Brown (1957) discuss these matters more fully, but the above is sufficient for the purposes herein.

We recognize at this point that the processes leading to the particle size distribution, and the lognormality of ϵ, are similar to Equation 2. In the latter case we need merely to replace Y_1 by $(1+Y_1)$. This adds a constant to the error term, but makes no fundamental change in the process.

A Simple Model of Pollutant Concentrations

It has been found that

$$\psi(\nu) = KQ/u(\nu) \tag{3}$$

where ψ is pollutant surface concentration
Q is source strength
u is wind speed
ν is frequency

---

*Approximation error small as $\Delta t \to 0$.

is an appropriate model for predicting concentration of SO_2 and particulates in a well mixed urban environment (Gifford and Hanna (1972)). The constant K has been measured rather extensively and a range constructed for K for each pollutant.

Based upon the lognormality of 1/u, which has been verified empirically (Knox and Lange (1972)), and the relative invariance of K, which has been said to be a weak function of city size, we conclude that ψ is lognormally distributed.

Knox and Lange (1972) have estimated $K' = KQ$ by experiment and by using a box model to predict concentrations. Their findings indicate that a suitable value of K can be found either by using the box model or comparing ψ and 1/u by visual superposition, and adjusting K' so that ψ and K'/u have approximately equivalent geometric means. In addition, with this value the variances of ψ and K'/u are approximately equal. (See Figs. 2-6 of Knox and Lange (1972)).

For continuous point sources Knox and Lange (1972) fitted the model

$$\frac{X(\nu)}{Q} = K_2 \frac{1}{\sigma_a(\nu)\,\sigma_e(\nu)\,U(\nu)} \tag{4}$$

where σ_a, σ_e are lateral and vertical standard deviations of the plume for a 5-year, argon-41 stack release at Chalk River, Ontario.

It was found that the frequency distribution predicted by this model, and the observed distribution, diverged significantly at higher concentration levels. This suggests that the relationship between $\psi(\nu)$ and $U(\nu)$ is more complicated than Equation 4 indicates. The fit to the lognormal distribution was also significantly poorer than that of the area sources. (See Fig. 9 of Knox and Lange (1972)).

We shall reconsider this point later in light of the findings presented below.

The General Model of Pollutant Concentration

The simple model, Equation 3, gives an indication of the fundamental reason for lognormality for well mixed urban areas where diffusion and photochemical terms are neglected. We can extend this argument to include the latter features by examining the differential equation predicting the time evolution of pollutant concentrations:

$$\frac{d\psi_a}{dt} + u\frac{\partial\psi_a}{\partial x} + v\frac{\partial\psi_a}{\partial y} + w\frac{\partial\psi_a}{\partial z} = \tag{5}$$

$$\frac{\partial}{\partial x}\left(K_x\frac{\partial\psi_a}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_y\frac{\partial\psi_a}{\partial y}\right)+\frac{\partial}{\partial y}\left(K_z\frac{\partial\psi_a}{\partial z}\right)$$

$$+\frac{Sa(x,y,z,t)}{V}+P(\psi_a,\psi_b\text{ ----- }\psi_n,t)$$

where ψ_a is the concentration of pollutant a; u, w, v are the velocity components; K_x, K_y are the lateral vertical eddy diffusitives, which are lognormally distributed based upon the lognormality of ϵ and the reproductive properties; K_z is the vertical eddy diffusivity; Sa is the source term for pollutant a; P is the term representing changes in concentration due to photochemistry; V is the volume of air for which S and P act.

This equation can be manipulated to represent a box model formulation (MacCracken, et al. (1972)) where we are concerned with the concentration averaged over a box which is surrounded by M other boxes.

$$\frac{d\left[\psi_k(m,t)\right]}{dt}=-\sum_{j=0}^{M}\left[T_A(m,j)+T_D(m,j)\right]\psi_k(m,t) \qquad (6)$$

$$+\sum_{j=0}^{M}\left[T_A(j,m)+T_D(j,m)\right]\psi_k(j,t)+S_k(m,t)$$

$$+P_k\left[\psi_a(m,t)\ldots\psi_n(m,t),t\right]$$

where T_A (m,j) and T_D (m,j) are the advective and eddy diffusive transfer coefficients from box m to box j. The lognormal distribution can be argued for these latter variables in a similar manner as for K_x and K_y.

This equation is also consistent with the generating process, Equation 2, when certain reasonable restrictions hold:

(a) The contribution of advection and diffusion terms are larger than the contribution of the source term. It has been found empirically that if this is not the case, lognormality does not result (Hopper (1972)).

(b) The concentrations in the surrounding boxes are, on the average over long periods of time, close to that of the box we are interested in, due to the fact that they are subjected to similar stimuli.

These restrictions transform Equation 4 to

$$\frac{d\psi(m,t)}{dt} = -\sum_{j=0}^{M}\left[T_A(m,j) + T_D(m,j)\right]\psi_k(m,t) \qquad (7)$$

$$+\sum_{j=0}^{M}\left[T_A(j,m) + T_D(j,m)\right]\psi_k(j,t)$$

Suppose we let

$$\psi_k(j,t) = \psi_k(m,t) + E_k(j,t) \qquad (8)$$

the equation becomes

$$\frac{d\psi(m,t)}{dt} = -\sum_{j=0}^{M}\left[T_A(m,j) + T_D(m,j)\right]\psi_k(m,t) \qquad (9)$$

$$+\sum_{j=0}^{M}\left[T_A(j,m) + T_D(j,m)\right]\psi_k(m,t) +$$

$$+\sum_{j=0}^{M}\left[T_A(j,m) + T_D(j,m)\right]E_k(j,t)$$

When we sum both sides in order to show lognormality, we have for the third term,

$$\sum_{T_O}^{T_R}\frac{\sum_{j=0}^{M}\left[T_A(j,m) + T_D(j,m)\right]E_k(j,t)}{\psi(t-\Delta t)} \qquad (10)$$

From meteorological reasoning we note that if the constant term is large, indicating strong winds, the difference between $\psi(j,t)$ and $\psi(m,t)$ will be small. Hence the term tends to zero. Conversely, in the case where the error term $E_k(j,t)$ is large, the constant term will usually be small indicating light winds. Furthermore, in either case, or any combination of cases occurring between T_o and T_R, we can expect that the sign of the term will vary, implying that the positive and negative terms will cancel each other.

This argument implies that Equation 9 is consistent with the law of proportional effect.

The solution will be source-dominated only when the magnitude of the source terms is comparable to the magnitude of the current concentration. There is reason to believe (Hopper (1972)) that in such cases the concentrations will not be lognormally distributed, as the model indicates. This result has been noted in investigations of particle size distributions also (Blifford and Gillette (1971)).

This reasoning is most easily justified for a well mixed urban region. It is not clear that the lognormal distribution will fit as well for non-urban, poorly mixed areas. We do feel, however, that the characteristics of an area's topography and typical meteorology would have to be highly unusual for (10) to be so large that the lognormal distribution would fit poorly.

We have not yet discussed the Δt interval necessary for these results. We recognize that it must be sufficiently small not to obscure the generating process. If, for an extreme example, Δt was 6 months, we would not see the process defined by Equation 2 because the effect of ψ_{i-1} on ψ_i would have long since died out. Larsen's (1965) data is for 5-minute instantaneous readings. We accept this as an appropriate time scale for our purposes, based on the fact that meteorology certainly does not change enough in a 5-minute period to obscure the relevant correlations. Of course, the distribution of pollutant concentration remains unaltered.

When the data is averaged over other time periods within the realm of atmospheric motion, the averaging time acts as a filter which smooths out motions of a smaller time scale. This has the effect of allowing us to see only motion of a time scale comparable to the averaging time in the averaged data. Hence the process described by Equation 2 still holds for larger averaging times, but the T_A, T_D terms now represent motion of a larger scale. This results in lognormality over a large spectrum of averaging times.

We are presently investigating the magnitudes of the first order multiplicative autocorrelations for all averaging times. Preliminary results indicate that significant positive autocorrelation is present for averaging times up to at least 2 weeks. This lends credence to the assumption that Equation 2 acts over a large spectrum of averaging times.

Applications

Ambient air quality standards (AAQS) are set in terms of the number of times a concentration of a particular pollutant shall exceed a specified limit, averaged over a specified number of hours. For example, an 8-hour average of CO may not exceed 30 ppm more than once per year.

Concentration distributions must be calculated to compare ambient air quality with these types of standards. An air quality prediction has meaning only when the averaging time and level of confidence of the estimate are included. This requires knowledge of the concentration distributions. Thus, whether we are interested in prescribing standards, describing levels, real time monitoring or land use planning, knowledge of concentration distributions is indispensable.

The foregoing discussion indicates that there is an increasing amount of evidence supporting the contention that surface air pollutant concentration frequency distributions are lognormal. This evidence includes empirical investigative results, arguments regarding the relationship of meteorological variable distributions to pollutant frequency distributions from simple diffusion models, and deductions of the nature of the pollutant frequency distributions from considerations of the complete set of governing equations for a multiple box model of photochemical pollutants. The possible exceptions to lognormality of pollutant distributions have been indicated. However, it is now pertinent to explore the practical and research implications of large portions of air quality regions having pollutant distributions that are lognormal; significant implications include:

(a) The application of air quality simulation models to land-use planning assessments for consistency with AAQS or to the design of measures to achieve consistency with AAQS in growing areas should be expedited in principle.

(b) The validation tests of air quality simulation models should include the requirement that calculated pollutant frequency distributions, or key portions of those distributions, correspond to reality.

(c) Knowledge that the pollutant concentration distributions are lognormal, should eventually lead to simplifications in data acquisition by air monitoring networks and to the feasibility of real time control mechanisms.

Land Use Plan Assessment

Consider the future when a verified and acceptable numerical simulation model for air pollution exists. The question then is, how can such an acceptable numerical simulation model be employed in land-use plan assessments? Given a region of interest for planning purposes and a suite of pollutants of concern, one could examine, for instance, the frequency distribution of hourly-average values of surface air concentrations and identify the portion of the distribution which is equal to or greater than the ambient air quality standard involved. Conceptually, the days or episodes involved in that part of the distribution could be composited into mesoscale or regional weather types. The meteorological fields and air quality data from those days or episodes would constitute case studies for model calculations. In Figure 3, 3 weather types are illustrated, corresponding to high, moderate, and low levels of pollution. The solutions of such numerical modeling case studies would delineate a spatial distribution of the excess over ambient air quality standards in the region which might not necessarily be defined by the network of monitoring stations. From examining those excesses and their spatial distributions, one could determine the degree of control and a location of control necessary to remedy the excess. In principle, the same set of analytical steps could be applied to forecast emission zonings associated with either growth or alternative land-use plans for the same set of identifed days. Hence, in this matter, one could evaluate the degree of control necessary for an existing situation in a region of interest to bring the air quality of that region into line with ambient air quality standards, or else to assess the excess of ambient air quality standards in need of control that correspond to various land-use plans.

Kennedy, et al. (1971) developed such a program for Chicago utilizing a sub-model to predict the effects of a certain type of emission zone in a particular place. These "coupling coefficients" are essentially a linear model of dispersion. They are used as coefficients in a linear program, the objective of which is to minimize the social and financial burden of restrictions while satisfying air quality constraints. Of course, non-linearities caused by interaction between pollutants and such are overlooked, and extremes are calculated through the use of coupling coefficients and extrapolation of the frequency distribution. But the model seems appropriate for making a land-use plan assessment or corresponding emission zoning.

Model Validation

Application of the model to such economically sensitive problems as land use planning requires that the model predict the surface concentration distributions quite accurately. In order to discuss validation of numerical simulation models of regional air pollution we reference some recent results in the development and initial verification of an air pollutant model for the San Francisco Bay Area (MacCracken, et al. (1972); Gelinas (1972)). This model uses historical meteorological data to predict the mean and surface air concentration in each of the model cells, including transport and diffusion by the ambient wind field between the irregular earth surface and the time and space variable marine inversion layer. (See Figs. 3-4 of MacCracken, et al. (1972)). The verification work was carried out on a 48-hour test period during July, 1968. Figure 4 displays the observed hourly-average concentrations of CO in parts per million during the case study, as well as the computed vertical average and computed surface hourly-average CO concentration. There is very reasonable agreement between the observed and the computed surface concentrations. This information of calculated versus observed concentrations can also be displayed as a lognormal frequency distribution plot, Figure 5. The significant feature to be noted here is that the frequency distribution of the predicted hourly-average concentrations on lognormal paper parallels the observed (Knox and Lange (1972)). In addition, it is parallel to that obtained by Larsen for the frequency distribution of hourly averages of CO for a year. Frank Gifford (ARATDL-NOAA) has recently noted that several of the numerical simulation models under development at this time render numerical solutions which are "noisier" than the observed distributions. A numerical solution that is contaminated with noise will, in general, not be able to predict the frequency distribution of the surface pollutant and, therefore, will have severe limitations in regard to a comparison of predicted frequency distributions to ambient air quality standards. Hence, one criterion for an acceptable model for numerical simulation of air pollution is whether the model is able to reproduce the frequency distribution characteristics of the pollutants involved and in the region of interest.

Monitoring

Knowledge of the particular distribution and its parameters allows us to make statistical comparisons between predicted air quality and air quality standards. Alternatively, we may simplify the procedure by taking random samples and manipulating only this reduced volume of data. The resulting estimates would be measures of typical long-term concentrations and variability. Figure 6 shows estimates of the distribution of hourly averages of CO in San Francisco from 1968 through 1970. The good agreement between estimates made from 100 random samples and 10 random samples, with the distribution obtained from continuous monitoring suggests the possibility that an appropriate spatial and temporal random sampling scheme would allow one movable receptor to estimate annual averages in a number of locations. This method has potential for use with land use models where long-term information is desired. Methods of sampling local air quality, as contrasted to continuous monitoring of local air quality, are not well suited to comparison of predicted concentrations from a model to short term AAQS.

Nonparametric methods were also investigated, but they tend to be less powerful than parametric methods in cases where the assumptions of parametric statistics apply. The latter methods also have the advantage of ease of manipulation and the simplicity of exact specification of the distribution.

A natural extension of the principles of air pollution monitoring is real-time control. This is a potentially effective method for controlling air pollution episodes. It requires a model with the ability to predict future pollutant concentration distributions at all points in the region sufficiently far in advance so that control actions can be taken to avoid an impending episode. These actions may be quite selective, in that they need only be taken during emergencies and then only in offending emission zones. We recognize that this is not within present capabilities, but we look ahead to the construction of such "feed forward" control schemes.

Conclusions

There is increasing evidence to support the theory that air pollutant concentrations are lognormally distributed in areas devoid of strong sources, whether they be passive or photochemical, and in the absence of meteorological or topographic effects resulting in sharp differences in concentrations between adjacent volumes of air. This lognormal distribution is supported by (a) empirical evidence, (b) the simple model of urban pollutant concentrations proposed by Gifford when examined in the light of the lognormal distribution of the reciprocal of wind speed verified by Knox and Lange, and (c) the theoretical derivation from the full set of equations governing the time evolution of pollutant concentrations presented herein.

Weibull wrote ". . . it is utterly hopeless to expect a theoretical basis for distribution functions such as . . . particle sizes," and yet one has been provided. In fact, it seems reasonable to expect that the physics describing a process should be consistent with a distribution function describing the results of that process, indeed, anything else would be suspect. This is what has been provided here, a consolidation of empirical evidence with physical theory.

Knowledge of pollutant concentration distributions is necessary for land-use plan assessment to compare predicted air quality with ambient air quality standards. It is useful as a method of verification of a numerical simulation model of air pollutant evolution, and it is a potentially valuable tool for use in a real-time model predicting short-term fluctuations in pollutant concentrations.

Acknowledgement

This work was performed under the auspices of the U. S. Atomic Energy Commission.

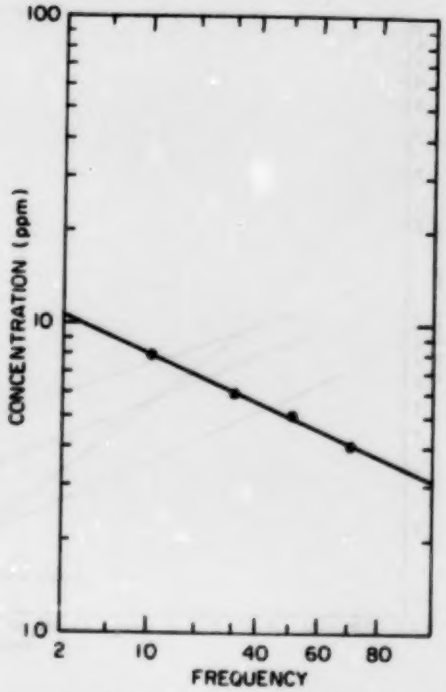


Figure 9-2. Lognormal probability plot of CO concentration vs frequency for San Francisco.

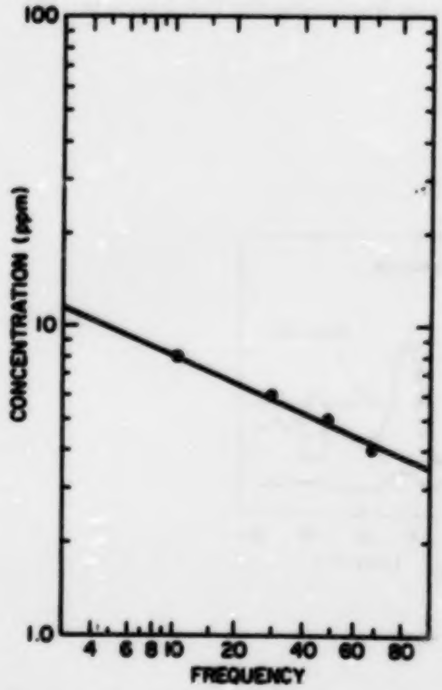


Figure 9-1. Weibull probability plot of CO concentration vs frequency for San Francisco.

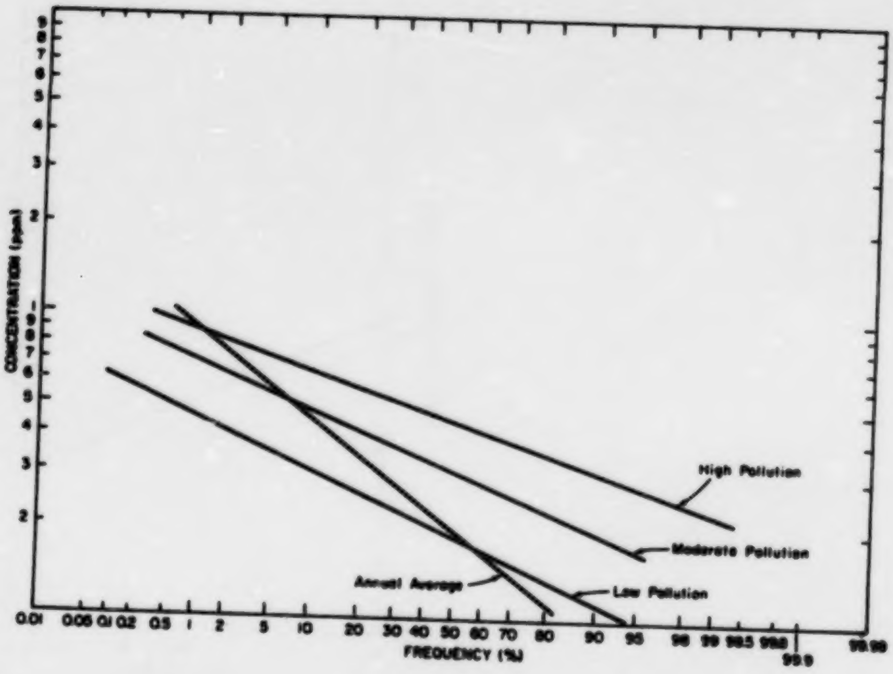


Figure 9-3. Carbon monoxide concentration vs frequency for San Francisco for various categories of pollution-days.

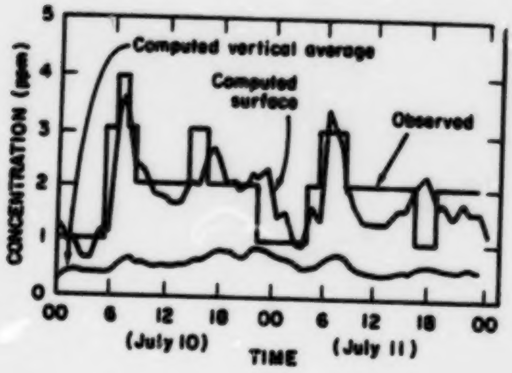


Figure 9-4. Carbon monoxide concentrations for San Francisco, July 10-11, 1968.

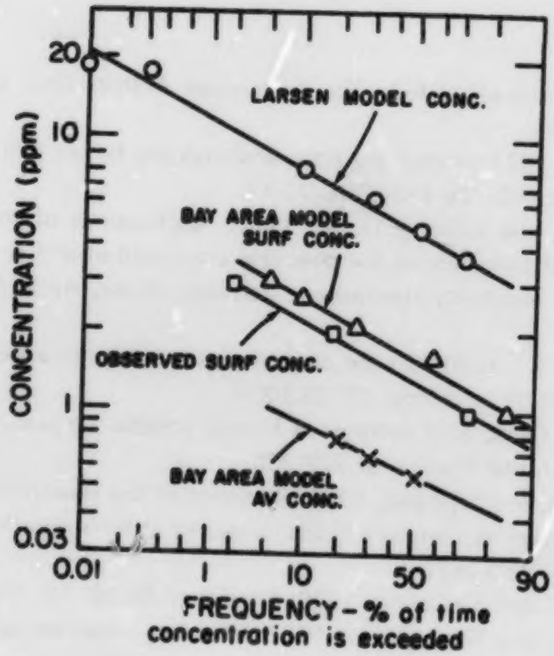


Figure 9-5. Carbon monoxide concentration vs frequency for San Francisco.

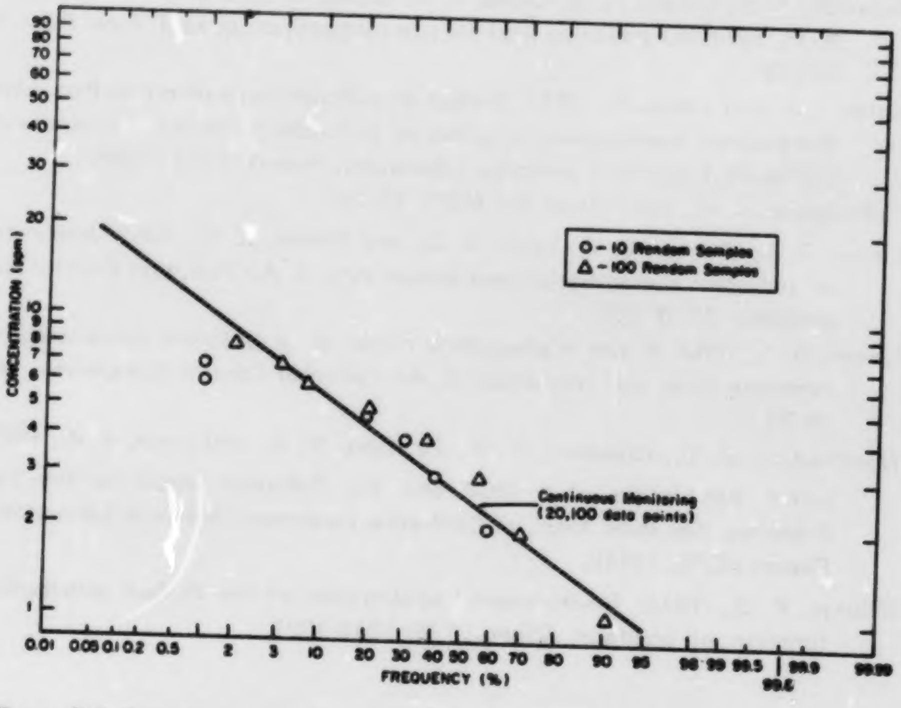


Figure 9-6. Carbon monoxide vs frequency for San Francisco—hourly averages.

References

Aitchison and Brown, 1957: *The Lognormal Distribution.* Cambridge Univ. Press, 176.

Barlow, R. E., 1971: Averaging time and maxima for air pollution concentration. NTIS AD-729 413, ORC 71-17.

Blifford, I. H., and Gillette, D. A., 1971: Applications of the lognormal frequency distribution to the chemical composition and size distribution of naturally occurring atmospheric aerosols. *Water, Air & Soil Pollution. 1:* 106-114.

Friedlander, S. K., 1960: *On the particle size spectrum of atmospheric aerosols.* Journal of Meteorology. 17: 373-374.

Gelinas, R. J., 1972: Stiff systems of kinetic equation, a practioner's view. *J. of Computational Physics. 9:* 222-236.

Gibson, Stegen, and Williams, 1970: Statistics of the fine structure of turbulent velocity and temperature fields measured at high reynolds number. *J. Fl. Mech. 41:* 153-167.

Gibson, Stegen, and McConnel, 1970: *Physics of Fluids. 13:* No. 10.

Gifford, F. A., and Hanna, S. R., 1972: Modeling urban air pollution. ARATDL Contribution No. 63.

Hopper, C., 1972: personal communication.

Kennedy, A. S., Cohen, A. S., Croke, F. J., Croke, K. G., Stork, J., and Hurter, A. P., 1971: Air pollution-land use planning project, phase I. Final Report, ANL/ES-7.

Knox, J. B. and Lange, R., 1972: Surface air pollutant concentration-frequency distribution: implications for urban air pollution modelling. University of California, Lawrence Livermore Laboratory, Report UCRL-73887.

Kolmogorov, A. N., 1941: *Dokl AN SSSR. 30:* 301.

Larsen, R. I., Zimmer, C. E., Lynn, D. A., and Blemel, K. G., 1967: Analyzing air pollutant concentration and dosage data. *J. Air Pollution Control Association. 17:* 85-93.

Larsen, R. I., 1969: A new mathematical model of air pollutant concentration, averaging time, and frequency. *J. Air Pollution Control Association. 19:* 24-30.

MacCracken, M. C., Crawford, T. V., Peterson, K. R., and Knox, J. B., 1972: Initial Application of a Multi-Box Air Pollution Model to the San Francisco Bay Area. Univ. of California, Lawrence Livermore Laboratory, Report UCRL-73348.

Milokaj, P. G., 1972: Environmental applications of the Weibull distribution function: oil pollution. *Science. 176:* 1019-1021.

Singpurwalla, N. D., 1972: Extreme values from a lognormal law with applications to air pollution problems. *Technometrics. 14:* 3.

Weibull, W., 1951: A distribution function of wide applicability. *J. Appl. Mech.* 293-297.

Zimmer. C. E., and Larsen, R. I., 1965: Calculating air quality and its control. *J. Air Pollution Control Association. 15:* 565-572.

DISCUSSION

J. Arvesen: Regarding your model itself, the box model that you applied to the San Francisco data, how do you go about estimating the parameters in that model to fit that data? Is there a problem involved with that? It would seem to be a problem to me. There seemed to be a lot of parameters in there and I was wondering how you can estimate them reasonably well on 2 days data. Am I missing something?

Knox: Let me see if I can answer the question. The predicted frequency distribution for the 48-hour test period was generated from the predicted 48 1-hour average CO concentrations for San Francisco receptor from the model. This distribution was compared to the actual data from San Francisco—the 48 average hourly values at the sampling station. And so the frequency associated with the highest CO value corresponds to 1 in 48. There is an interesting aspect of this: the obvious question is how do we know that the model has an averaging time that is appropriate to be compared to the average hourly data. If one looks at the boxes used, they are "T" shaped, "L" shaped, or any arbitrary shape that fits the area roughness or source strength. Their average dimension divided by wind velocity is about an hour, so that the travel time across the boxes is comparable to sampling period. If we had used 5-minute integrations, then the comparison to actual data should be performed with 5-minute average CO data.

10. AIR QUALITY FREQUENCY DISTRIBUTIONS FROM DISPERSION MODELS COMPARED WITH MEASUREMENTS

D. BRUCE TURNER*

Environmental Protection Agency
National Environmental Research Center
Division of Meteorology
Research Triangle Park, North Carolina

Introduction

Cumulative frequency distributions (hereinafter abbreviated CFD) of air quality can be estimated by dispersion models. By comparison with CFD's from air quality measurements at the same location, some indication of the accuracy of these estimates can be made. Extremes of the estimated CFD for specific locations can be compared with air quality standards. Not only can estimates be made for existing pollution sources, but projected estimates can be made for expected degrees of control of existing sources and inclusion of additional sources. These projected estimates can also be compared with air quality standards.

It is the purpose of this paper to present CFD's estimated from short-term dispersion models and determined from measurements for the same locations, periods of record and averaging times, and to compare these, especially the maximum value, to indicate the accuracy of the estimates.

Background

National ambient air quality standards have been set in response to the Clean Air Act. In most cases the standards consist of a long-term average, usually the annual average, and a short-term standard, such as a maximum 24-hour or 3-hour concentration not to be exceeded more than once per year. For existing sources, it is possible to monitor ambient air quality at selected sites to determine if air quality standards are met at these locations. Due to the small number of monitoring stations, it is highly likely that maximum concentrations occur that exceed those measured at these stations.

---

*On Assignment from the National Oceanic and Atmospheric Administration, Department of Commerce

It is desirable to supplement present air quality measurements by estimating concentrations at additional locations. It is also desirable to estimate projected ambient air quality at a number of locations for proposed source configurations including both dditional sources and various assumptions as to degree of control. These estimates can also be compared with air quality standards. Air quality dispersion models have been developed to meet this need.

Long-term or climatological models have been used to estimate mean annual concentrations at specific locations. These models typically require mean annual emission rates from point and area sources and joint frequency distributions of wind direction, wind speed, and stability. The relative accuracy of these models is discussed elsewhere (Turner, Zimmerman and Busse (1972)). Summarizing this paper, comparison of model estimates with measurements at a number of sampling locations indicates that the ratio of root mean square error to the measured mean for all stations is typically from 0.3 to 0.5. This indicates that annual means can be estimated quite well. These estimated means can be compared with the standards for the annual mean.

Dispersion models that calculate concentrations for averaging times of 1 to 2 hours can be used to make estimates for comparison with short-term standards. Calculations can be made for each hour of the period of record and, in addition to determining the extreme concentration occuring once during the period, a frequency distribution of concentrations can be obtained. Hourly concentrations can also be averaged for any longer averaging time, such as 24 hours, and a frequency distribution determined for this longer averaging time. These short-term dispersion models require both meteorological and emission information. Meteorological information typically consists of (a) wind speed and direction or wind flow fields, and (b) atmospheric stability class and mixing height or temperature variation with height. Emission information typically consists of emission rates for both significant point sources and all other sources considered collectively as area sources. To be realistic the variations in emissions from season to season, weekday to weekend and for various times of the day should be included. It has been the experience of the author that this information is difficult to obtain and also difficult to organize into a convenient form. Stack parameter data are usually included for the point sources in order to calculate plume rise. Because of inclusion of most emissions near the ground into area sources, the resulting concentration estimates represent concentrations averaged over an area the size of the smallest area source, usually 1 km^2. On the other hand, air quality measurements represent the concentration at the specific point of measurement and are therefore particularly sensitive to any nearby sources. Validation of dispersion models in urban areas is therefore difficult, since it is necessary to compare the point measurements with estimates that are more representative of an area.

Frequency Distributions From Dispersion Models

Fortak (1970) and Koch and Thayer (1972) have estimated CFD's for locations in urban areas from short-term dispersion models. Both used Gaussian plume models, making separate calculations for point and area sources.

Fortak had 30-minute measurements of sulfur dioxide for four locations in the city of Bremen, Germany. He made estimates using short-term dispersion models for the same locations and averaging times and determined the frequency distributions over various periods. The following results are for the heating period (20 September 1967 - 31 May 1968). At two stations estimates are higher than measurements for corresponding percentiles throughout the distribution. At another station, estimates are less than measurements over the entire distribution. For the remaining station, estimates are higher than measurements except beyond the 99.6 percentile of the CFD where estimates are too low. At the extreme end of the CFD, at the 99.5 percentile, Fortak's estimates for all four stations are well within a factor of 2 of the measurements. The worst estimate is off by a factor of 1.7.

Koch and Thayer (1972) of Geomet, working on a contract for EPA, also used a short-term dispersion model to estimate 1-hour concentrations for 8 locations in Chicago for a 1-month period, (January 1967), and to estimate 2-hour concentrations for 10 locations in St. Louis for a 3-month period (December 1964 - February 1965). CFD's were determined from these estimates and compared with CFD's from measurements at the same locations.

In Chicago, the model underestimates concentrations for the entire CFD at one of the stations. Four stations have concentrations overestimated for the entire CFD. At one of the stations, concentrations are overestimated at the low end of the CFD with slight underestimates past the 55 percentile. The other two stations have concentrations underestimated at the low end of the CFD and overestimated beyond the 65 percentile for one station and beyond about the 90 percentile for the other. Only one station has an estimate at the 90 percentile off by more than a factor of 2. The error at this station is a factor of 2.8. For these CFD's the 90 percentile is the highest cumulative frequency for which data is presented.

In St. Louis, the model underestimates concentrations for the entire CFD for five of the stations. One station has concentrations overestimated for the entire distribution. At the other four stations concentrations are generally underestimated, but are overestimated at the top end of the CFD with the cross-over ranging from the 55 percentile to the 90 percentile. Only one station has an estimate at the 90 percentile off by more than a factor of 2 from the measurement at the same point in the CFD. The error at this station is a factor of 2.6.

24-Hour Frequency Distributions

The author, using a short-term Gaussian plume model similar to that used by Koch and Thayer (1972), calculated 2-hour concentrations for 40 locations in St. Louis. Measurements of 24-hour sulfur dioxide concentrations were made at these stations during 89 consecutive days in December, 1964 through February, 1965. Estimates of 24-hour concentrations at these stations were made by averaging 12 successive 2-hour estimates. Frequency distributions of 24-hour concentrations for the period were determined for both estimated concentrations and for measured concentrations for all 40 stations.

Because of the interest in the extreme end of the CFD (at the frequency of the air quality standards), the extreme estimated value and the extreme measured value were compared. These are near the 99 percentile for these three months of 24-hour concentrations. The ratio of calculated concentration to observed concentration was determined for each station. These ratios for the extreme, arranged in ascending order, are 0.63, 0.70, 0.78, 0.81, 0.82, 0.84, 0.84, 0.87, 0.88, 0.88,0.90, 0.97, 1.07, 1.07, 1.09, 1.10, 1.11, 1.20, 1.23, 1.24, 1.30, 1.32, 1.42, 1.42, 1.44, 1.45, 1.45, 1.46, 1.59, 1.63, 1.65, 1.67, 1.69, 1.79, 1.90, 2.05, 2.23, 2.34, 2.35, 2.37.

Note that 35 of the 40 stations have estimated extreme values within a factor of 2 of the measured extreme (ratio between 0.5 and 2.0). Also 15 stations have errors of less than or equal to ±20%.

Examples of agreement of estimates from the model and measurements at the extreme (around the 99 percentile) are shown in Figures 1 through 3. Station 4 (Fig. 1) has the best agreement (a ratio of 0.97). Station 23 (Fig. 2) has the highest overestimate (off by a factor of 2.37). Station 27 (Fig. 3) has the greatest underestimate (a ratio of 0.63).

The comparison of the CFD's for the 40 locations is characterized subjectively as follows: At ten stations the CFD's for estimates and measurements are close. At ten stations overestimates occur throughout the entire CFD. At three stations underestimates occur throughout the entire distribution. At four stations overestimates occur primarily, but underestimates occur at the higher percentiles (beyond the 88, 93, 95, and 96 percentiles). At 10 stations both underestimates and overestimates occur, with overestimates beyond the crossover points of 7, 10, 25, 25, 25, 25, 30, 40, 83, and 90 percentiles. At two stations, although both underestimates and overestimates occured, the comparison could be described as mostly underestimates. At one station underestimation occured except at each end of the distribution.

Other visual comparisons of the estimated and measured CFD's for 24-hour concentrations are given in Figures 4 and 5. Station 8 (Fig. 4) has the best agreement between estimates and measurements over the whole CFD and has a ratio of 1.07 at the extreme. Station 23 (Fig. 2), discussed previously, has the worst overestimate irregardless of place in the distribution, with the estimate four times the measurement at the 7 percentile. Station 38 (Fig. 5) has the worst underestimate with an estimated 2 and a measured 46 at the 4 percentile, off by a factor of 23. This is probably because low levels of background concentration exist due to emissions from distant sources that are not included in the calculations made by the model.

It is also desirable to consider if the CFD's appear to be lognormal (straight lines on log-probability plots), particularly in view of the frequent use of the Larsen statistical model (Larsen (1971)) to estimate extremes of concentrations in urban areas. It appears that there is some deviation from the lognormal distribution in the figures previously discussed, especially Stations 27 (Fig. 3) and 38 (Fig. 5). Stations 28 (Fig. 6) and 16 (Fig. 7) seem to have two different slopes in their distributions of measured concentrations, with the transition taking place in the vicinity of the 50 percentile. Stations 2 (Fig. 8) and 6 (Fig. 9) have a sudden transition to higher measured concentrations around the 95 to 97 percentiles. Station 19 (Fig. 10) has two portions of the CFD of measured concentrations with the same slope but with a displacement occurring near the 50 percentile. For the most part, measured concentration CFD's appear to be near lognormal. Although many of the CFD's from estimated concentrations are also nearly lognormal, some of them appear to deviate more than those of the measurements and to have an "S" shape such as station 28 (Fig. 6).

Two-Hour Frequency Distributions

At 10 of the 40 measurement stations in St. Louis, 2-hour measurements of sulfur dioxide were also made. At these 10 locations, estimates and measurements were used to determine CFD's for 2-hour concentrations over the 89 day period (December 1964 - February 1965).

For each station, the extreme estimated value and the extreme measured value were compared. Since the data period consisted of 12 periods per day for 89 days, the extreme represents a frequency near the 99.9 percentile. The ratio of calculated concentration to observed concentration was determined for each station. These ratios for the extreme, arranged in ascending order, are 0.52, 0.66, 0.74, 0.75, 1.12, 1.47, 1.51, 1.60, 1.76, 1.88. All 10 of the stations have the estimated extreme value within a factor of 2 of the measured extreme (ratio between 0.5 and 2.0).

A selected number of these 2-hour CFD's are shown in Figures 11 through 13. Station 17 (Fig. 11) has the best agreement at the 99.9 percentile. Station 36 (Fig. 12) has the highest overestimate at the 99.9 percentile (off by a factor of 1.88). Station 10 (Fig. 13) has the greatest underestimate at the 99.9 percentile (a ratio of 0.52).

The comparison of the 2-hour CFD's for the ten locations is characterized subjectively as follows: Two stations (4 and 12) have cumulative frequency distributions close to those of the measurements. At two stations (3 and 23) overestimates of concentration occur for the entire CFD with the largest errors less than a factor of 3. At two stations (10 and 33) underestimates of concentration occur for the entire distribution with errors as large as a factor of 4. At three of the stations (17, 28 and 36) concentrations are overcalculated beyond the following percentiles: 99, 56, and 63. At one station (15) concentrations are undercalculated beyond the 45 percentile.

Other visual comparisons of the estimated and measured CFD's for 2-hour concentrations are given in Figures 14 through 16. Station 4 (Fig. 14) has the best agreement between estimates and measurements over the whole distribution. Stations 23 and 28 (Fig. 15 and 16) have poor agreement between estimates and measurements throughout most of the CFD. At station 23 concentrations are primarily overestimated. At station 28 concentrations are underestimated at low percentiles and overestimated at high percentiles.

The two measured CFD's that are least lognormal occur at stations 33 and 36. Station 33 (Fig. 17) appears to have two slopes, and at the highest concentration (greater than 99.8 percentile) there is a sudden increase in concentration. Station 36 (Fig. 12) also seems to have two different slopes with the transition occurring around the 70 percentile. For estimated CFD's, station 28, 33, and 36 (Figs. 16, 17, and 12) appear to be least lognormal.

Discussion

These CFD's from measured air quality data and from dispersion model estimates have been determined for averaging times from 30 minutes to 24 hours, for periods of record from 1 month to a heating season. These are all for locations within urban areas. These cannot be compared directly to present U. S. air quality standards since the standards specify periods of record of 1 year. However, it is quite likely that during the heating season in Bremen, and during December through February in St. Louis, the highest sulfur dioxide concentration of the year occurs, due to the number of space heating sources that produce sulfur dioxide. Concentrations with the extreme frequency of once per year should be expected to vary considerably from year to year, due to the high variability of occurrence of stagnant or other special meteorological conditions that cause the extreme.

The number of stations with extreme estimates from the dispersion models within a factor of 2 of the extreme measurements for the investigators mentioned herein are summarized in Table I.

Dr. Frank Pasquill (1971) in his presidential address delivered before the Royal Meteorological Society on April 21, 1971 stated, "The agreement as close as 20 or 30 percent which may be achievable in the most favorable circumstances for a long-term multi-station average, is obviously unattainable in respect of an individual value even when this is averaged over an hour or so. In this case the only prospect of useful prediction lies in the statistics of the cumulative frequency distribution of a large number of such values, and it would appear. . . that prediction of the rather extreme high concentrations encountered only occasionally may be achievable with an error factor of about two."

Since most of the extreme value estimates are within a factor of two of the extreme measurements, these results are in agreement with Dr. Pasquill's statement.

It should be pointed out that these model results mentioned here contain both overestimates and underestimates so that no constant correction factor can be used to bring the estimate of these extremes more in line with the measured extremes. Errors in both directions with regard to emissions and small sources near the receptor probably account for a large proportion of the differences. Keep in mind that model estimates representative of areas a square kilometer or larger are being compared with point measurements.

There are many other comparisons and statistical tests than can be performed with these CFD's in addition to the consideration of the extremes and the rather cursory examination of the lognormality of them. Some of the possibilities for further examination of this data follow: Perform statistical tests to determine how close the given CFD's are to lognormal. Determine standard geometric deviations (slope of distribution) from two percentiles in the distribution and see how these vary with location in the urban area. From the 40-station sampling network determine measured and estimated concentration patterns at various percentile levels. Determine what meteorological conditions cause the extreme value estimated concentrations and the extreme value measured concentrations at each station.

Conclusions

Gaussian plume dispersion models for urban areas produce CFD's at individual sampling locations similar to the distributions determined from measurements. These distributions subjectively appear similar to lognormal distributions. The maximum 24-hour concentration estimated during an 89-day period was within a factor of 2 of the measured maximum at 35 of 40 sampling stations in St. Louis, Missouri. The maximum 2-hour concentration estimated during the same 89-day period was within a factor of 2 of the measured 2-hour maximum at all 10 sampling stations, having 2-hour measurements available. Estimates of air quality concentration at a downwind receptor for a given hour from a point source are generally regarded as accurate only within a factor of 2 because of uncertainties in estimates of emission rate, turbulence structure, plume height, wind direction and wind speed. It is encouraging to find similar accuracies for the extreme value (99 percentile for 24-hour, 99.9 percentile for 2-hour) estimates for urban locations influenced by multiple sources. (Note that the maximum estimate may be calculated for a different

2-hour period than the period that has the maximum measured concentration.) This gives somewhat increased confidence to the air pollution meteorologist asked to estimate urban air quality concentrations to be compared with standards. One must keep in mind that good estimates of concentrations from dispersion models can only result from good emission estimates and reliable measurements of meteorological parameters.

Acknowledgements

The author wishes to thank Adrian D. Busse for his development some years ago of a computer program to routinely produce a cumulative frequency distribution from a time series of data, Dale H. Coventry for programming the computer-plotter routine to produce log-probability plots, Ralph I. Larsen for suggesting the preparation of this paper, and Lea Prince for her valuable assistance.

Table I

Number of stations with extreme estimates from models within a factor of two of extreme measurements, and worst error.

Investigator	Fortak	Koch and Thayer		Turner	
City Averaging Time	Bremen 30-min.	Chicago 1-hour	St. Louis 2-hour	St. Louis 2-hour	St. Louis 24-hour
Extreme Percentile	99.5	90	90	99.9	99
Within a factor of 2	4 of 4	7 of 8	9 of 10	10 of 10	35 of 40
Worst error, a factor of:	1.7	2.8	2.6	1.9	2.4

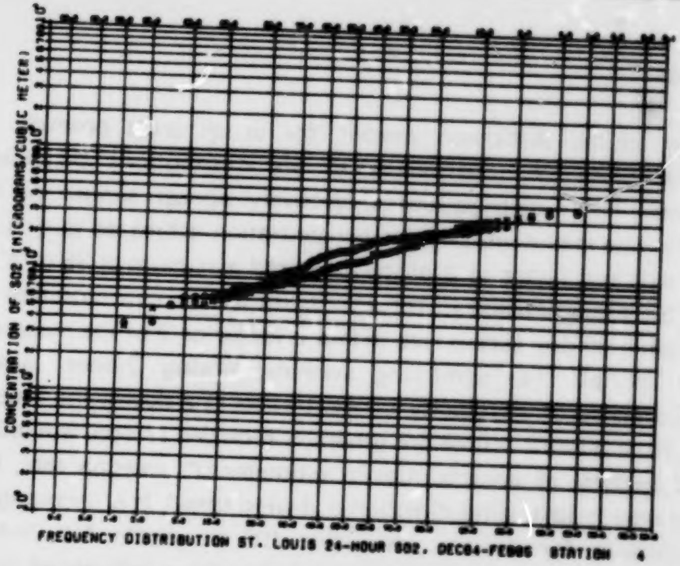


Figure 10-1. Best agreement at the 99th percentile.

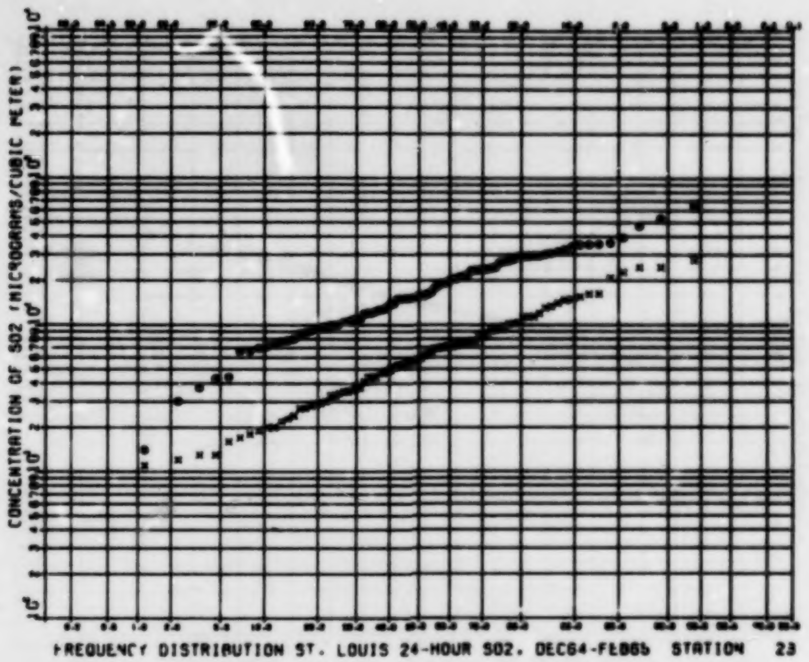


Figure 10-2. Highest overestimate at the 99th percentile.

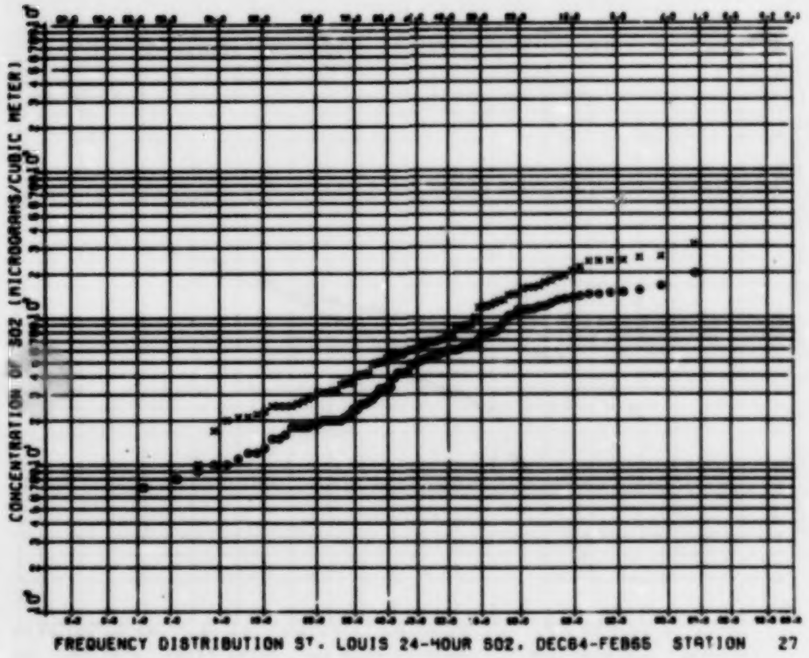


Figure 10-3. Greatest underestimate at the 99th percentile.

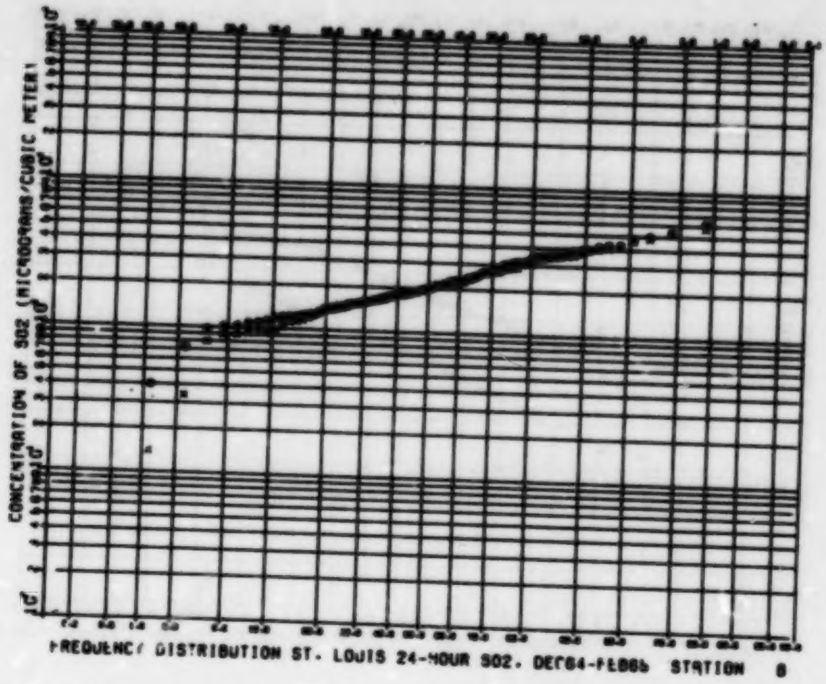


Figure 10-4. Best agreement over the whole cumulative frequency distribution.

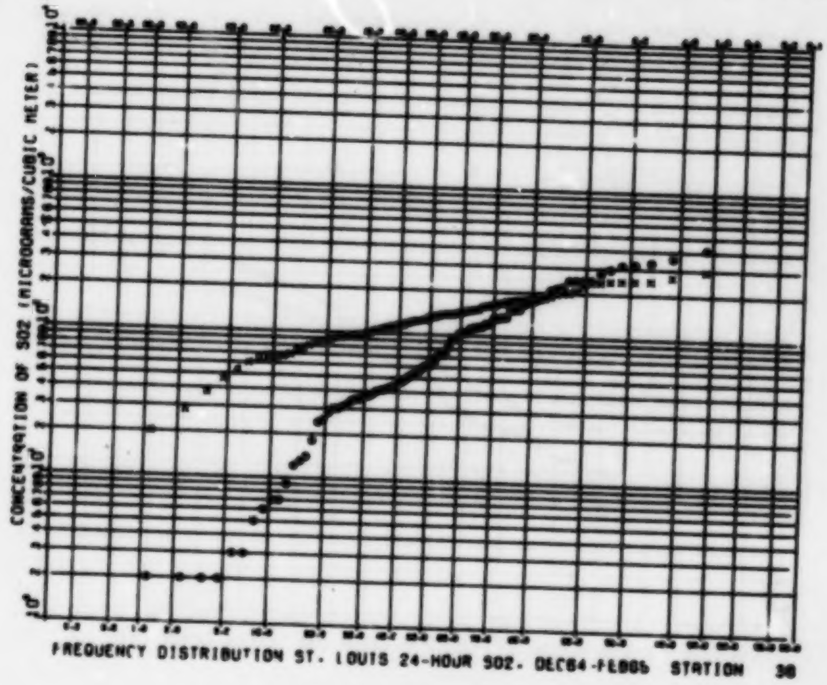


Figure 10-5. Greatest underestimate.

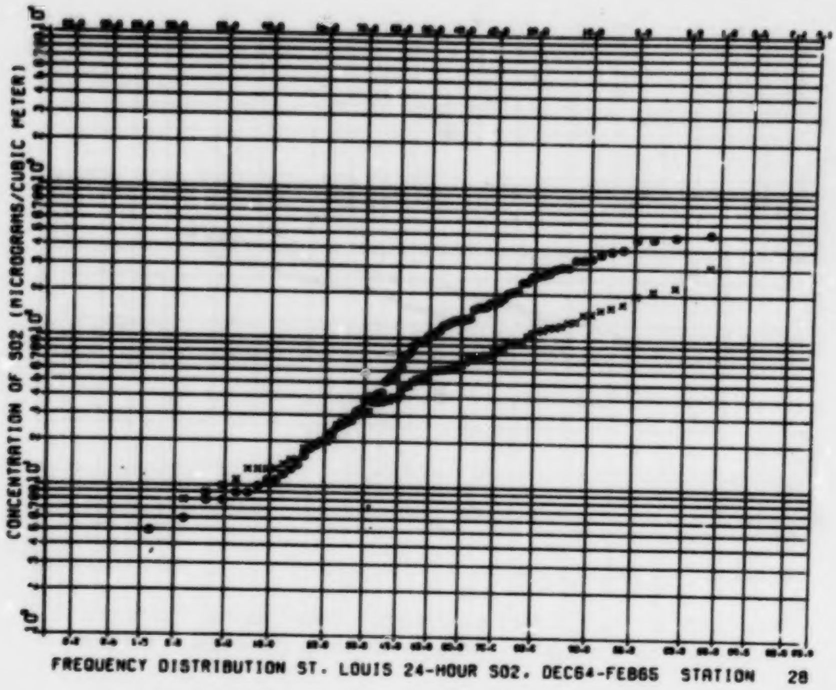


Figure 10-6. Example of two different slopes.

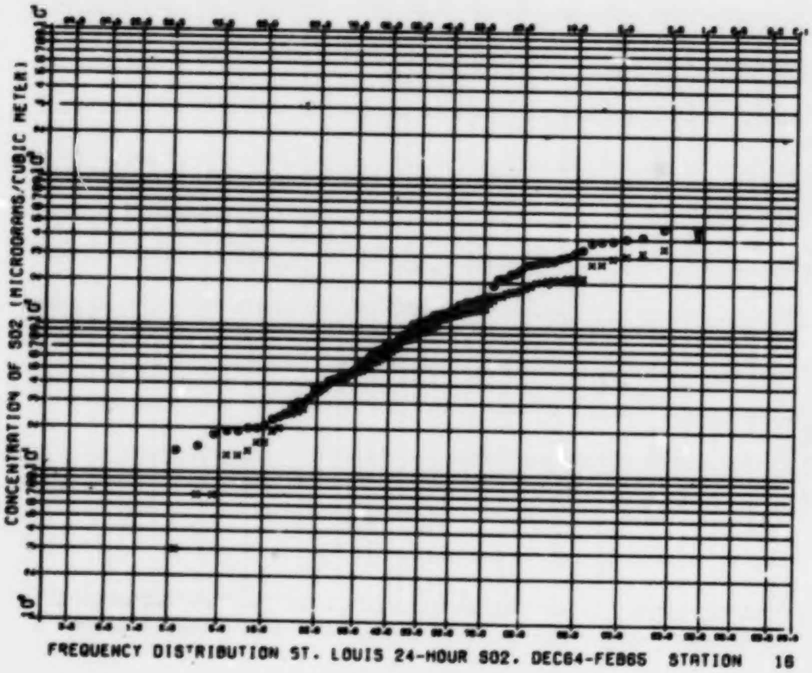


Figure 10-7. Example of two different slopes.

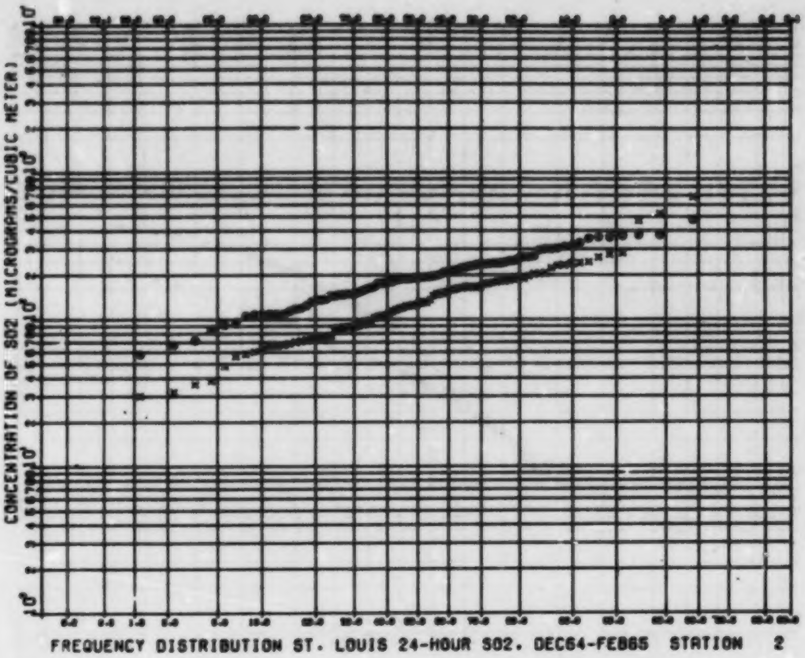


Figure 10-8. Example of sudden transition to higher measured concentrations.

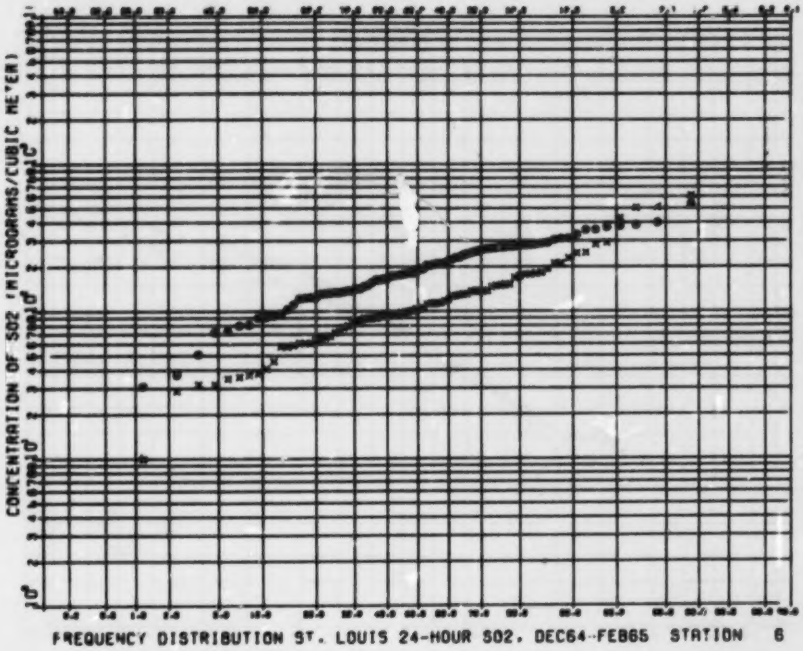


Figure 10-9. Example of sudden transition to higher measured concentrations.

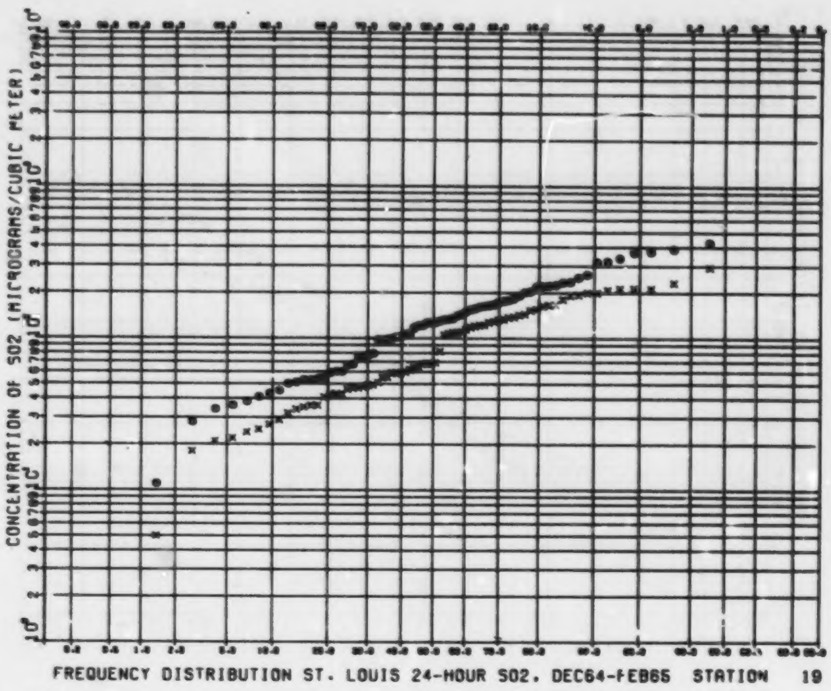


Figure 10-10. Example of sudden transition (both portions have same slope).

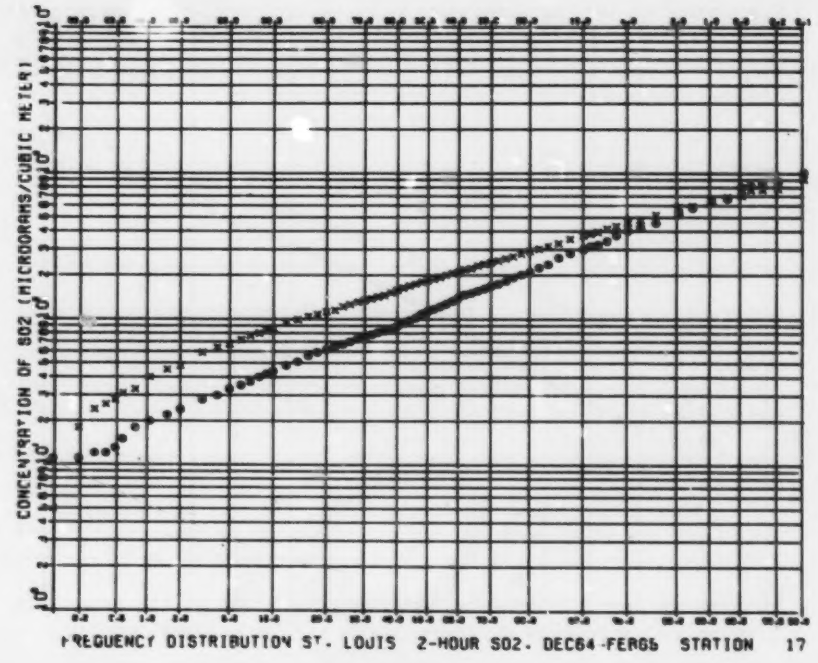


Figure 10-11. Best agreement at the 99.9 percentile.

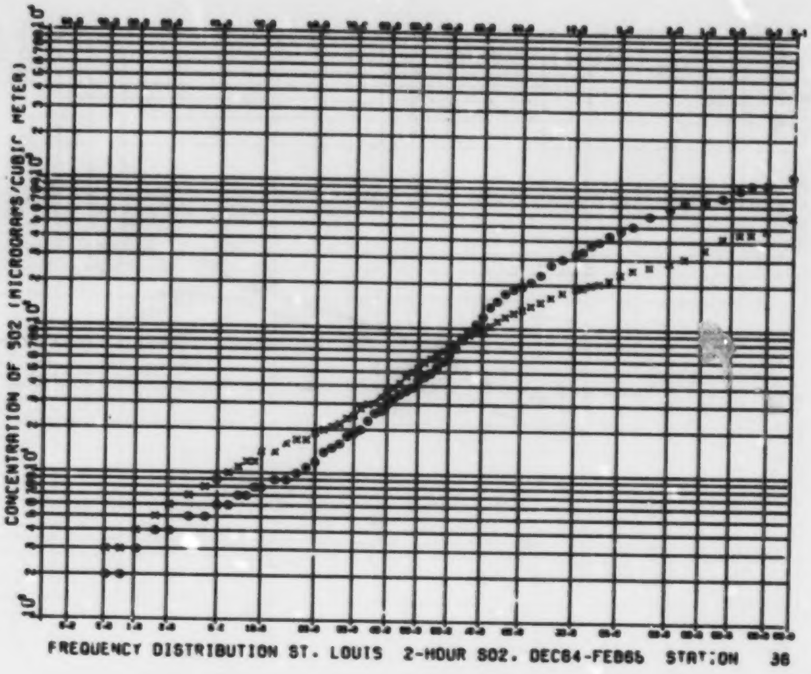


Figure 10-12. Highest overestimates at the 99.9 percentile.

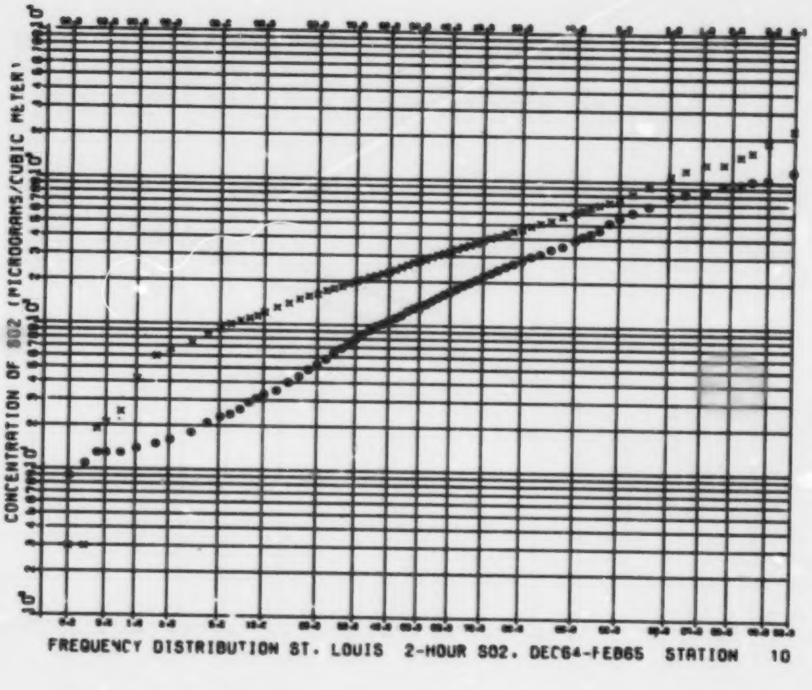


Figure 10-13. Greatest underestimate at the 99.9 percentile.

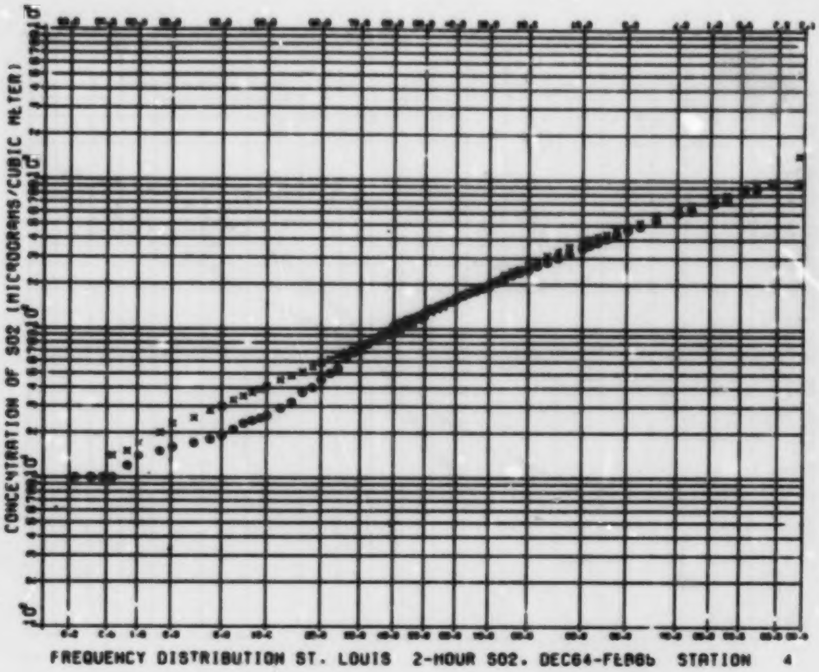


Figure 10-14. Best agreement over whole cumulative frequency distribution.

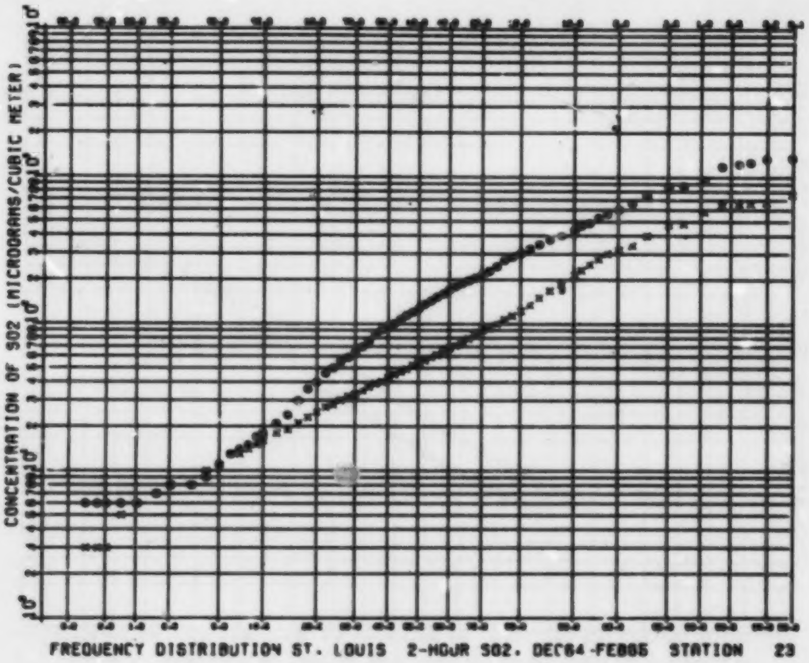


Figure 10-15. Poor agreement, primarily overestimation.

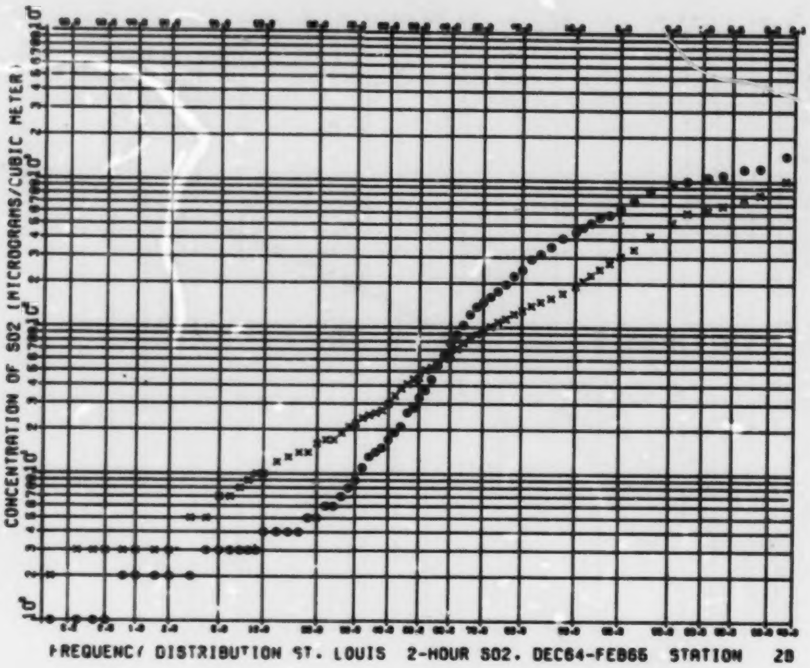


Figure 10-16. Poor agreement, underestimates and overestimates.

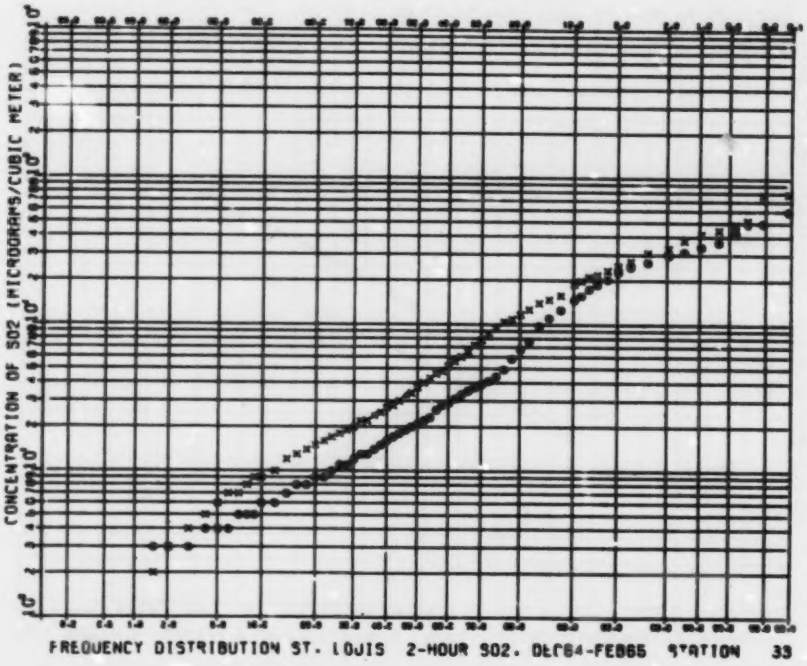


Figure 10-17. Example of sudden transition to higher measured concentrations.

References

Fortak, G., 1970: Numerical simulation of temporal and spatial distributions of urban air pollution concentration. Proceedings of Symposium on Multiple-Source Urban Diffusion Models. EPA, Air Pollution Control Office Publication No. AP-86.

Koch, R. C., and Thayer, S. D., 1972: Validation and sensitivity analysis of the Gaussian plume multiple-source urban diffusion model. Final Report prepared under Contract Number CPA 70-94, Geomet, Inc. EPA Office of Air Programs Publication No. APTD-0935.

Larsen, R. I., 1971: A mathematical model for relating air quality measurements to air quality standards. Environmental Protection Agency, Office of Air Programs Publication No. AP-89.

Pasquill, F., 1971: Atmospheric dispersion of pollution. *Quarterly J. Royal Meteorol. Soc. 97:* 369-395.

Turner, D. B., Zimmerman, J. R. and Busse, A. D., 1972: An evaluation of some climatological dispersion models. Presented at 3rd meeting of Panel on Modeling of NATO Committee on the Challenges of Modern Society, Paris, France.

DISCUSSION

J. Visalli: It has been suggested in some papers that were presented earlier that, particularly for SO_2, the body is susceptible to even shorter-term fluctuations than 2 hours; I'm talking about fluctuations more on the order of 2 to 4 minutes. I was wondering how good you feel your model would be in predicting variations for this short of a time interval.

Turner: I would like to have good representative 2- to 4-minute meteorological information and good emission information that includes the variability of all sources with the time interval of 2 to 4 minutes, in order to attempt such short-term concentration estimates.

Singer: One comment touched on a bit before by Don Pack and also questioned by Frank Gifford, and a slight comment which you made at the end of your paper, pleased me. Everyone has been dealing with numbers and just using them blindly without making any comment about the accuracy of the data and bringing in the statistics. The source term, Q, many times is out by an order of magnitude when you actually look at the data itself. The meteorology may be out by a factor of 2. When you start verifying it and looking at the SO_2 data, which can also easily be out. I would like to see someone say what is the accuracy of the data and try to bring that into the statistics, the meteorology and also the final verification. You said you were out by 20, this could easily just simply be an instrument error. But I would like to see this aspect of statistics. I mean, Don asked that question this morning—can you bring in the error into the analysis or can you verify it. I have always heard the answer; yes (it can be done) but I have never seen anyone do it.

Turner: With regard to the accuracy of the data Irv, I have to give the people credit who did the sampling in St. Louis for their care to obtain good data. For 24-hour samples they did have replicate bubblers side by side. Air was drawn through them by the same pump but with two different critical orifices, one for each of the samplers. These duplicate samples were compared in the laboratory. I forget the exact numbers, but on the order of 3 to 5 percent of the total number of samples deviated by more than 10 percent. Most of these were thrown out for the reason that they did not duplicate within 10%.

Singer: I knew your data was good, but it was just a general warning to the statisticians who take the number that we provide and blindly use it. We know better in that respect.

J. Rustagi: I also notice this kind of behavior is lognormal. Actually, there are outlier-prone distributions. I don't know whether lognormal is one of them but Professor Neyman has given a detailed analysis of outlier proneness of distributions in a symposium which was conducted at Columbus in 1971.* The gamma distribution is one of them. This is one approach which could be taken. Secondly, as has been mentioned before, that if the concentration is too low or too high we have different kinds of errors of measurement. Suppose that you use the same model as lognormal and you have the variance dependent on the mean. In my data it was noticed that at low levels the errors were proportional to that of the mean. So if you put in the model the variance as a certain function of the mean, the estimate of mean and so on can be correspondingly calculated.

Helmut Lieth: I can verify your statement for the low values from our analysis of the national air pollution network data, but what do you do with the variability in the high levels?

Rustagi: As you said if there are different kinds of behavior at lower end and upper end one could put the variance as a quadratic function of the mean, cubic function of the mean, some other function of the mean, or some other complicated function. What I mean is that the variance should reflect the error in measurement as noticed by instrumentalists.

Lieth: Yes, but there is a problem. There is a logical difference in the production of the high values and subsequently their variability, and the low values. It is a factual problem as well that you have more of some kind of pollutant at a certain weather condition. So this is not plainly a statistical deviance. How can you get this logical difference out with a straightforward statistical method?

D. McNeil: We attempted to look at that problem with some data in New Jersey and the point of view that we adopted was to try and find a transformation of the pollutant concentration which would make the variance constant. In fact, that was how we arrived at the fourth root transformation. In doing so we rather

---

*J. S. Hustagi (Editor), ***Symposium on Optimizing Methods in Statistics,*** Academic Press, New York, N. Y., 1971.

fortuitously found it made the mean value of the increment in the level linear as the function of time. That would be one way of solving that problem. You can do that . . . you might find you need a different transformation depending on the weather conditions, but we did it just by lumping all the values for one year together.

Don Pack: I think we could belabor this to death, but I would like to point out one thing that when one is dealing in trying to identify extreme values you want the longest possible period of record. I did a little arithmetic back there. New York is generating around three million two hundred thousand estimates of an individual pollutant for about ten pollutants. Alright, you've got thirty million values, very attractive to people with large computers. However, the error function in a real measurement system is not stationary. It trends with time. Initially, and I point specifically to Mr. Turner's data—research data carefully controlled with operators who were dedicated to producing the best possible information. On the other hand, the kind of data that is becoming available to us in the many urban areas of order 10 to 20 cities are not of that kind at all. The technicians may be very devoted initially and the equipment will be new. But with time everything deteriorates so that we would have error functions such that as the length of your record increases, errors also increase. The only point that I am trying to make is that deductions on the kinds of distributions can be markedly affected by the character of data.

L. Crow: In studying extreme values and measurements of particulates in a natural background in Wyoming, some important meteorological influences have been noted. Natural dust produces the very high extremes due to high winds, but the high winds are not neatly distributed throughout this year or any other year. The extremely low values are affected by precipitation. Wind blown dust can be locked-in during winter by heavy snow cover. Is there a way that we as meteorologists and statisticians can treat these extreme ends using real data instead of some arbitrary formula? Can we bring about an adjustment for the extremes that do occur if we add the meteorological parameters to actual instances of extreme data?

Rustagi: There would be a way of mixing distributions if we know enough about the distribution at the other extremes. There are procedures available for estimating parameters with mixtures and you can put the distributions in two different tails with corresponding probabilities—that would be one possible way.

Lieth: I think we have listed in the paper that we handed out here a little while ago, the new program NONREG, which probably solves that problem mathematically for you. NONREG is in a package available here on the UNC campus.

Court: I can't help being impressed by the great similarity of the problems being discussed today and those with which we have been dealing for many years in the field of hydro-meteorology—such problems as the inaccuracies of rain gages, the non-normality of rainfall, the various procedures such as cube root and fourth root to obtain homoskedacicity for regressions and many other similar relations.

Neustadter: I would desire to talk with anyone who might be able to help me with a problem that we have almost at hand in our program. I mentioned very passingly at the beginning of my presentation that we have hundreds of samples which we're subjecting to analysis. These are samples collected on high-vols on high quality analysis paper, and we are doing a lot of analysis. Essentially what we are going to end up with is a set of hundred's of items each characterized by ten's of parameters. We have been looking for techniques and so far we don't know that much about it, but pattern recognition seems to sound like the best thing. The only thing we are aware of is one article from Livermore that seems to indicate that pattern recognition is now coming into the field of chemistry and handling multiple parameter chemical reactions and phase changes.

11. FOURIER ANALYSIS OF AIR MONITORING DATA

BERNARD E. SALTZMAN

Department of Environmental Health
University of Cincinnati
Cincinnati, Ohio

The proliferation of pollutant monitoring activities is now providing massive amounts of data. Effective utilization of this information requires proper analysis. This may be as costly as the collection of the data. A major problem has been to obtain the "signal" from the data in the presence of overwhelming amounts of "noise" produced by environmental fluctuations. Computers have been utilized to provide information on the statistical distributions of the numbers. Tabulations also have been presented of data averaged by time of day and/or by season (NAPCA (1969)). The purpose of this paper is to explore the application of another technique, Fourier analysis of data, which offers the promise of extracting significant new types of information.

In Figure 1, a plot is given of monitoring data for particulate lead in air (Cholak, et al. (1968)). This was obtained by continuously sampling outside air from the second floor window at the Kettering Laboratory, Cincinnati, Ohio, through 4-inch membrane filters. The filters were changed on Mondays, Wednesdays and Fridays; thus the lead analyses represented values averaged for 2- or 3-day-periods. Application of the usual statistical calculations provided the following information: mean 1.07 $\mu g/m^3$, standard deviation 0.55 $\mu g/m^3$, geometric mean 0.95 $\mu g/m^3$, standard geometric deviation 1.60. A plot of cumulative and differential frequency distributions is given in Figure 2, which shows a tendency to a lognormal distribution. What other significant information can be extracted from this data?

Examination of Figure 1 indicates irregular fluctuations with time. These can be regarded as analogous to colored light, comprised of the sum of a mean value and of a series of fluctuations of differing periods, amplitudes and phases. In the case of a mixture of colors of visible light, resolution into a spectrum can be obtained by the means of a spectroscope. In the case of sound or radio wave mixtures, tuned circuits can be utilized to obtain the spectra. Recent developments in computer science now make practical Fourier analysis of data, which is the equivalent of a spectroscope in providing the spectra of fluctuations. A good explanation of this technique for chemists has been presented by Horlick (1972). In order to explore these possibilities, computer programs were prepared utilizing a Wang computer and plotter, which was available and convenient for program development. Table I provides a summary of the programs that were developed.

Explanation of Program

Data for this program should consist of a series of values at uniform time intervals. The time units are usually hours or days. Provisions are made for missing data. The fluctuations are resolved as the sum of a series of sine and cosine waves of different amplitudes and periods. Thirty-six periods are used covering 7 octaves (doublings) from 3 times to 384 times the time interval of the data; each octave is divided into 5 equal, logarithmically-spaced steps.

Data Processing

For each data point, the time from the middle of its interval to a selected initial reference time and date (e.g., midnight on a Sunday) is calculated. This time is divided by the first period (3 time units), and converted to a time phase angle; the date value is multiplied by the sine of the angle and stored in one register, and by the cosine and stored in a second register. This calculation is repeated for successively longer periods (up to 384 time units), and the data stored in 70 other registers. About 10,000 computer steps are required for each data point. Each of the 72 data registers accumulates sine or cosine products for its assigned period. Mathematically, the calculations are as follows:

For each value of data X_j taken at time t_j, a series of 36 calculations is made by assigning to an index, i, consecutive integral values from 0 to 35:

$$\text{Period, } p_i = 3 \times 2^{i/5}$$

$$\text{Sine term, } S_{ij} = X_j \sin \frac{360\, t_j}{p_i}$$

$$\text{Cosine term, } C_{ij} = X_j \cos \frac{360\, t_j}{p_i}$$

The 36 pairs of registers are each assigned to a specific period, p_i They accumulate the corresponding sums $\Sigma_j S_{ij}$ and $\Sigma_j C_{ij}$ for all data points.

Data Printout

In the data printout, the final time, t, from the reference time to the end of the last data value is given. The number of time units of data, d, is tabulated. The mean value is calculated as follows:

$$\bar{x} = \Sigma x_j / d$$

Rather than presenting the sine and cosine results separately, a clearer picture is obtained by combining them into a vector sum and a phase angle. The latter is combined with the period to calculate the first peak time after the selected reference time.

For each period tabulated, the results are calculated as follows:

amplitude,

$$A_i = \frac{2}{d} \sqrt{\left[\sum_j s_{ij}\right]^2 + \left[\sum_j c_{ij}\right]^2}$$

The peak time, t_i, (past the reference time) is calculated as follows:

peak degrees,

$$\theta_i = \arctan \left(\frac{\sum_j s_{ij}}{\sum_j c_{ij}} \right)$$

peak time,

$$t_i = \left[\frac{\theta_i}{360} \right] \times p_i$$

If the fluctuations are in phase with the cosine wave (peak at reference time), the resulting angle, and peak time are zero.

Data Plotting

In both types of data plots, the horizontal scale is a logarithmic scale of periods. The initial point represents 3 time units, and each inch represents 1 octave (doubling of period). In the amplitude plot, the vertical scale above the origin is a linear scale, on which 5 inches is equal to the amplitude range selected. The plotted points present the spectrum of fluctuation intensities. In the peak time plot, the vertical scale below the origin is a linear time scale beginning at the time selected. Each 0.1 inch represents 6 time units. The scale can be marked off in appropriate divisions, e.g., days of the week or months of the year. The plotted points represent the first peak time and 6 consecutive subsequent ones. For periods exceeding 64 time units, fewer peaks are plotted because the maximum ordinate is 384 time units, or 6.4 inches downward.

Results

Table II illustrates the data printout of the computer program. This program required that the data values be entered as integers. The values, which were expressed in micrograms per cubic meter to the nearest hundredth, were therefore multiplied by 100 before entry. The first column gives the selected values of period, which are the same for all data processing. Each is approximately 1.149 times the previous one. Each fifth line represents exactly 1 octave, or a doubling of the period. It will be noted that due to slight inaccuracies in the computer, the values for 48, 192, and 384 are listed respectively as 47.999, 191.999, and 383.999. Amplitude values are given in the second column. It can be seen that there are several peak values. The third column presents the peak times as phase angles. Since these are not convenient to visualize, the fourth column presents the times for the first peak past the selected reference time. In this case, the units are days after January 1, 1967.

A clearer visualization of these results can be seen from the plot of the data in the first and second columns of Table II, shown in Figure 3. Surprisingly, there is a dip in the amplitude line at a period of 7 days, although there are peaks at 3 1/2, 6, and 8 days. There are also successive amplitude peaks at the following multiples of 8 days: 2, 4, 6, 8, 12. Since the Fourier calculations are not accurate unless the data cover a time interval of at least 4 periods, the plotted values for periods exceeding 100 days cannot be considered as accurate. These data represent 364 days of measurement.

The computer output also includes phase information. Figure 4 is a plot of values in the first and fourth columns of Table II. The vertical scale downward represents a linear peak time scale, which is marked off in months of the year. The horizontal scale is a logarithmic representation of cycle periods, identical with that of the horizontal scale in Figure 3. To understand the significance of this plot, one may visualize a straight line vertically downward for the period 96 days. It can be seen from column 4, Table II that the first peak time occurs 51 days after January 1st (or February 21st). This is indicated in Figure 4 by a dot towards the bottom of the box representing February. There are successive dots vertically downward for each 96 days thereafter. Thus the dots and the connecting lines represent the times during the year when each cycle maximum occurs. If Figure 4 is viewed vertically below Figure 3, peak times are shown in correspondence with each amplitude value plotted in Figure 3. To avoid crowding on the left side of the figure, for each period no more than 7 peaks are plotted. In this plot the vertical time scale begins at January 1, 1967. The computer program also permits starting this vertical scale at any desired time after the reference time. This is the equivalent of shifting the plotted lines and the time scale vertically upward and viewing any selected lower portion. The lower end can be understood to extend to infinite time.

In the preceding discussion it was indicated that data for many cycles of period were required for accurate results. Figure 5 presents Fourier spectral amplitude data for the 3-month period of October-December, 1968 for total hydrocarbons in Cincinnati. The data for this and the two following figures were hourly-averaged values reported (NAPCA (1969)) by the Continuous Air Monitoring Program of the National Air Pollution Control Administration. Surprisingly, again there is no peak at 7 days. Major amplitude peaks can be seen at 12 hrs., 18 hrs., and 1, 3 1/2, 6, 8, and 12 days. Similar plots were made for the hydrocarbon data for each individual month of October, November and December. The patterns of amplitude peaks showed a similarity although their proportions were altered for the different months. The combined data for the 3 months eliminated some of the erroneous high peaks that were obtained for periods exceeding 1 week. As the amount of data increases, the sharpness increases of the "tuning" of the calculations for each period, and some of the peaks are reduced.

Figure 6 shows the Fourier amplitude spectrum for sulfur dioxide concentrations in Cincinnati, for the month of October, 1968. Amplitude peaks are evident at periods of 12 and 18 hours, and 1, 4 1/2, 6, and 8 days. If this figure is compared with the hydrocarbon results in Figure 5, it can be seen that there is a remarkable similarity, even though these pollutants come from entirely different sources. This suggests that the atmospheric dispersion processes, which are similar for both pollutants, exert the major controlling role in determining the atmospheric levels of these pollutants. Figure 3 also shows peaks at corresponding periods. All of these figures show a dip at 7 days and peaks at 6 and 8 days. They all show evidences of peaks at 3 1/2 days.

Figure 7 shows the phase results for the sulfur dioxide data. The downward time scale in this case is from 0 to 16 days. The days of the week are indicated on it. The computer program can view any portion of these results, which can be assumed to extend downwards to infinity.

Discussion

The significance of the Fourier spectra presented will become clearer after more types of data from more locations are analyzed. Interesting possibilities are opened up by this technique. Common factors operating to determine pollutant levels should become evident by amplitude spectra peaks in alignment. Differing periods indicate differing sources of variation. The Fourier analysis technique also offers a means of correcting the data for the incomplete response characteristics of the sampling methods or of the instrumentation. It has been shown (Saltzman (1970); Schnelle and Neeley (1972); Horlich (1972)) that the resultant data include distortions because of failure to respond to rapid changes.

If the transient and frequency responses are known, they can be incorporated in the Fourier computer program. This may permit recalculation of the data to correct for the distortions and more closely approximate the actual levels in the atmosphere.

The calculations described above were carried out on a small computing system which was readily accessible and convenient to rapidly develop a program. The Wang system requires approximately 1 millisecond for each step. Approximately 10,000 steps were required for the calculations on each data point. Thus the calculations for hourly data for 1-month period (720 data points) required 100 minutes of computer time. A program is being developed for an IBM S/360/65 computer. Preliminary results are in agreement with those already presented. The IBM computer, of course, has a much greater capacity, and can calculate for more intervals of period, allowing finer detail or greater range. Calculating time was found to be 500 times as fast as that of the Wang system. Future work should show whether results in other cities are parallel to those in Cincinnati.

Acknowledgement

This work was supported in part by the Center for the Study of the Human Environment, under U. S. Public Health Service Grant ES00159, and in part by the Environmental Protection Agency Grant R800869.

Table I
Summary of Fourier Programs for
Wang Model 700B Computer
With Model 702 Plotting Output Writer

Name of Program	No. Blocks	Total No. of Steps	Functions
3-Digit Data Recording	2	718	Records, edits, and retrieves 3-digit numbers on magnetic tape cassette; 48 numbers in each block, 100 blocks on each side of cassette.
Fourier Analysis of Data	4	911	Retrieves 3-digit numbers from tape cassette, performs Fourier calculations, tabulates and plots results.
Recording and Retrieval of Fourier Data	1	352	Records on magnetic tape cassette and retrieves contents of 74 Fourier calculation registers before they are altered by data printout and plotting. Permits adding and subtracting of blocks of data.

Table II

Data Printout of Program

Cincinnati, Lead in Air ($\mu g/m^3 \times 100$)

2nd Floor, Kettering Laboratory

1/13/67 to 12/31/67

Reference Time: Sunday, January 1, 1967

Mean 106.812, Data Time 352.00, Final Time 364.00

Period	Amplitude	Peak, Degrees	Peak Time
3.000	1.666	65.082	.542
3.446	2.550	181.068	1.733
3.958	2.000	58.181	.639
4.547	3.789	282.193	3.564
5.223	5.270	141.800	2.057
6.000	7.042	90.440	1.507
6.892	3.955	313.694	6.005
7.917	10.084	263.186	5.787
9.094	4.754	91.834	2.319
10.446	5.065	239.137	6.939
12.000	8.175	242.803	8.093
13.784	8.230	152.877	5.853
15.834	12.176	132.093	5.809
18.188	2.727	341.833	17.273
20.893	8.819	265.902	15.432
24.000	7.131	330.826	22.055
27.568	9.846	68.748	5.264
31.668	15.490	45.089	3.966
36.377	3.478	56.746	5.734
41.786	9.145	58.009	6.733
47.999	15.615	156.919	20.922
55.137	10.204	84.015	12.867
63.336	23.046	170.716	30.034
72.754	6.397	299.028	60.432
83.572	9.391	98.225	22.802
96.000	12.771	190.476	50.793
110.275	10.484	121.677	37.272
126.672	18.866	156.335	55.115
145.508	28.376	103.599	41.874
167.145	15.357	67.377	31.282
191.999	16.456	120.906	64.483
220.550	33.781	92.264	56.524
253.345	38.536	48.657	34.241
291.017	32.181	355.468	287.354
334.291	28.957	281.102	261.027
383.999	44.370	216.058	230.462

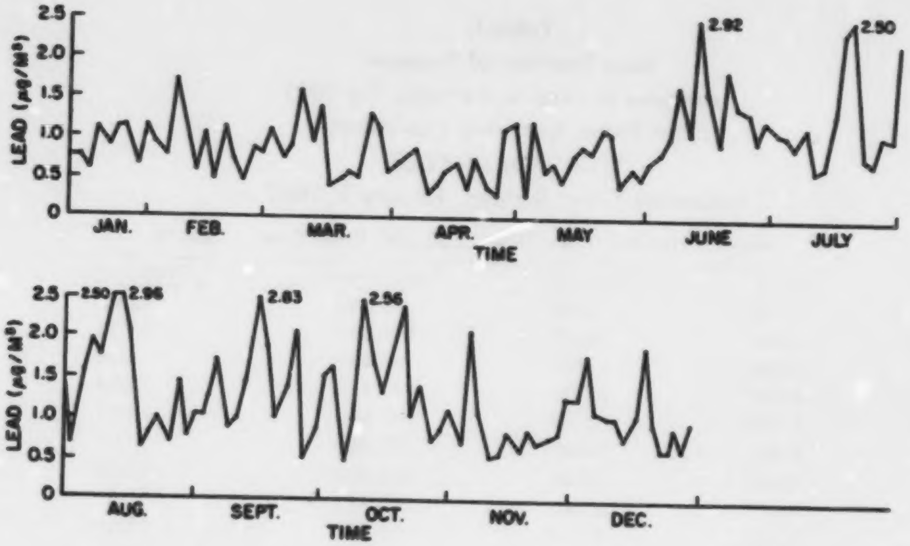


Figure 11-1. Concentrations of lead in air sampled from the second floor window at the Kettering Laboratory, Cincinnati, for the period January 1 to December 31, 1967.

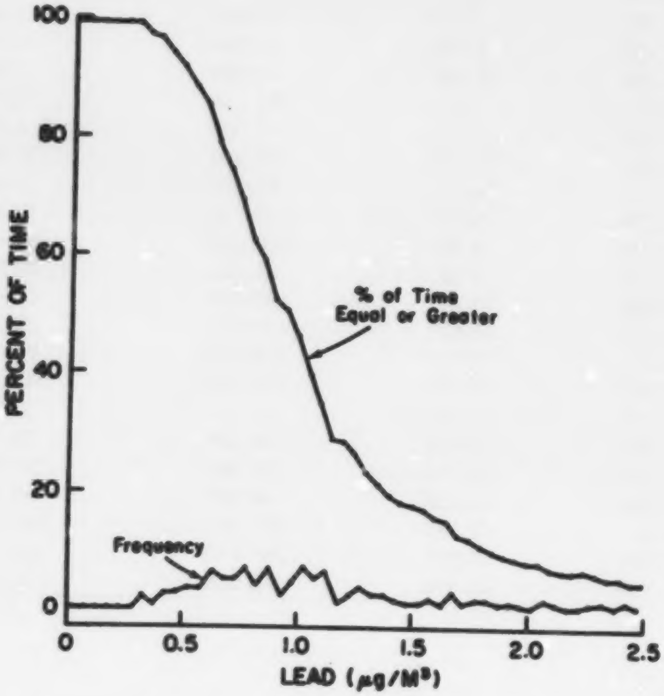


Figure 11-2. Differential and cumulative frequency distributions of the concentrations of lead in air shown in Figure 1.

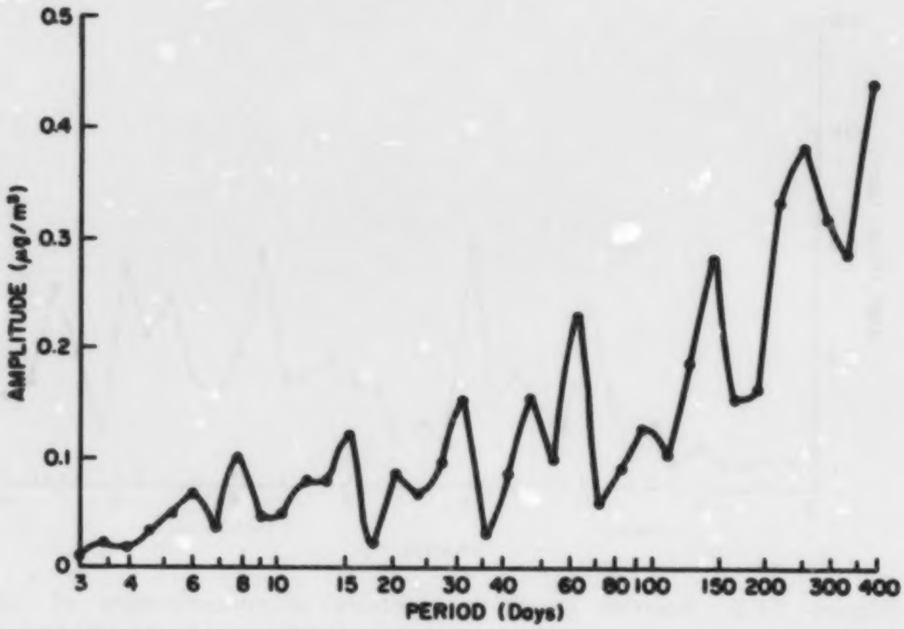


Figure 11-3. Fourier amplitude spectrum of the concentrations of lead in air shown in Figure 1.

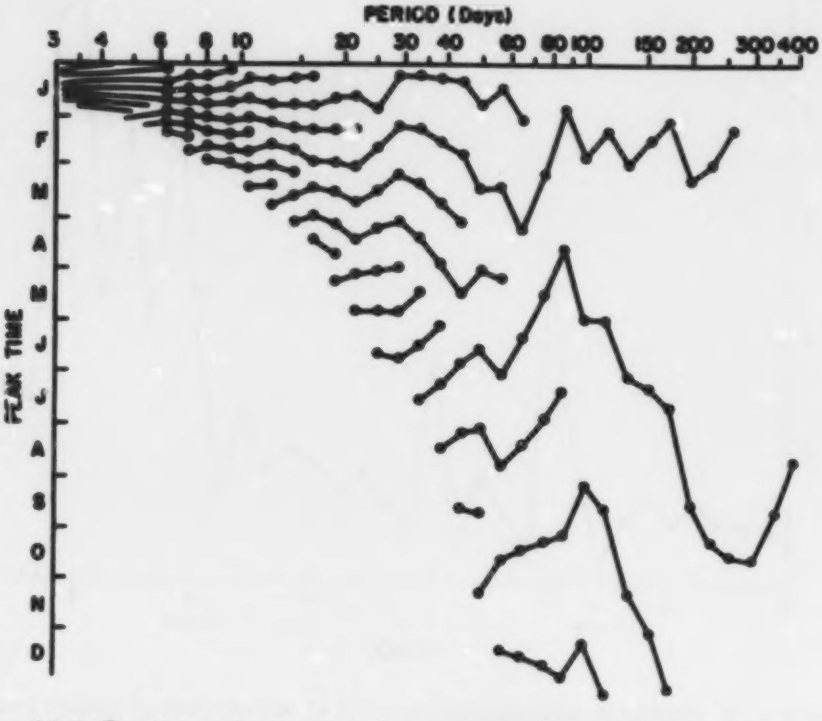


Figure 11-4. Fourier peak time data for concentrations of lead in air shown in Figure 1.

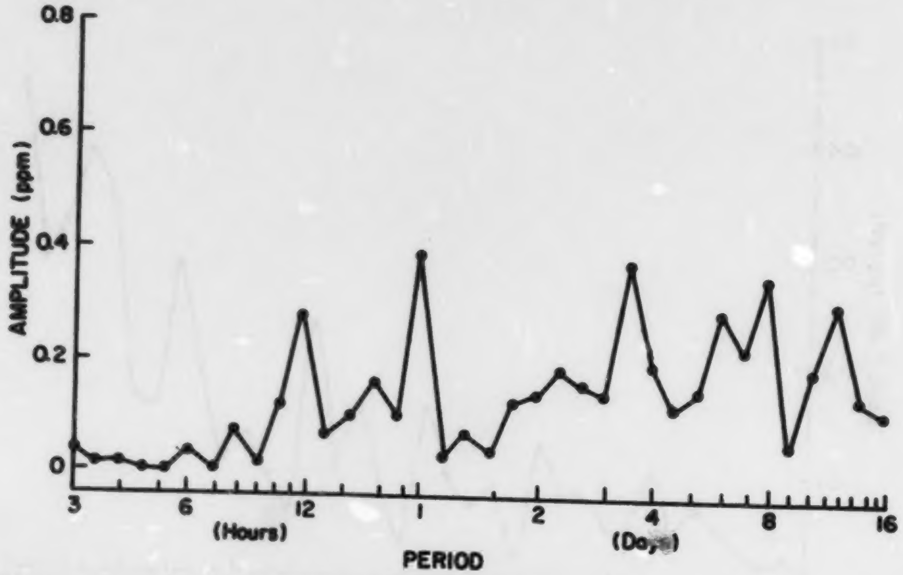


Figure 11-5. Fourier amplitude spectrum of concentrations of total hydrocarbons in Cincinnati for Oct.-Dec., 1968 (as reported by CAMP).

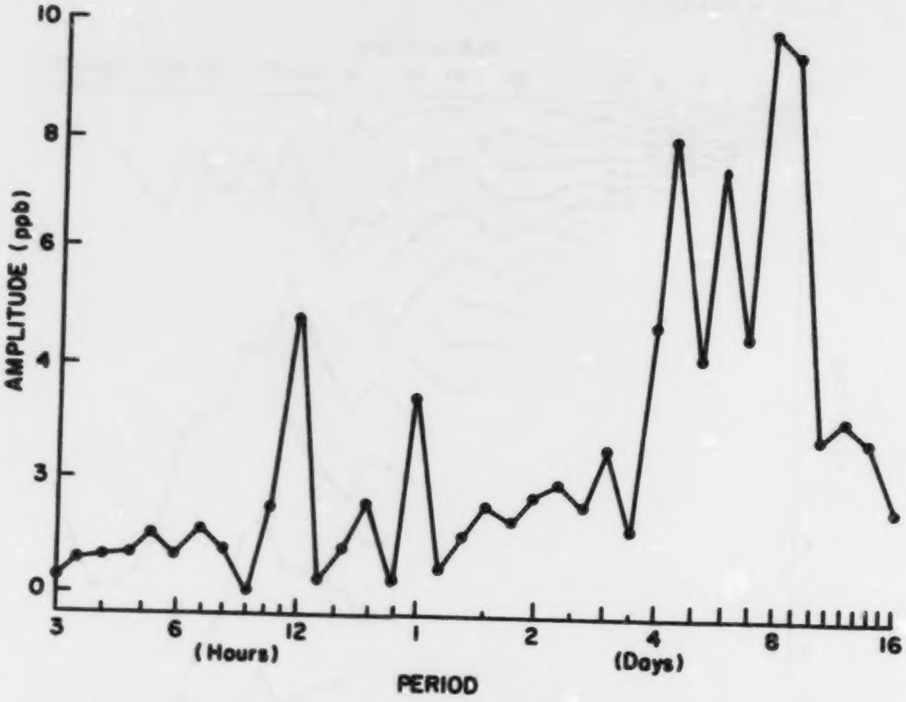


Figure 11-6. Fourier amplitude spectrum for concentrations of sulfur dioxide in air in Cincinnati for October 1968 (as reported by CAMP).

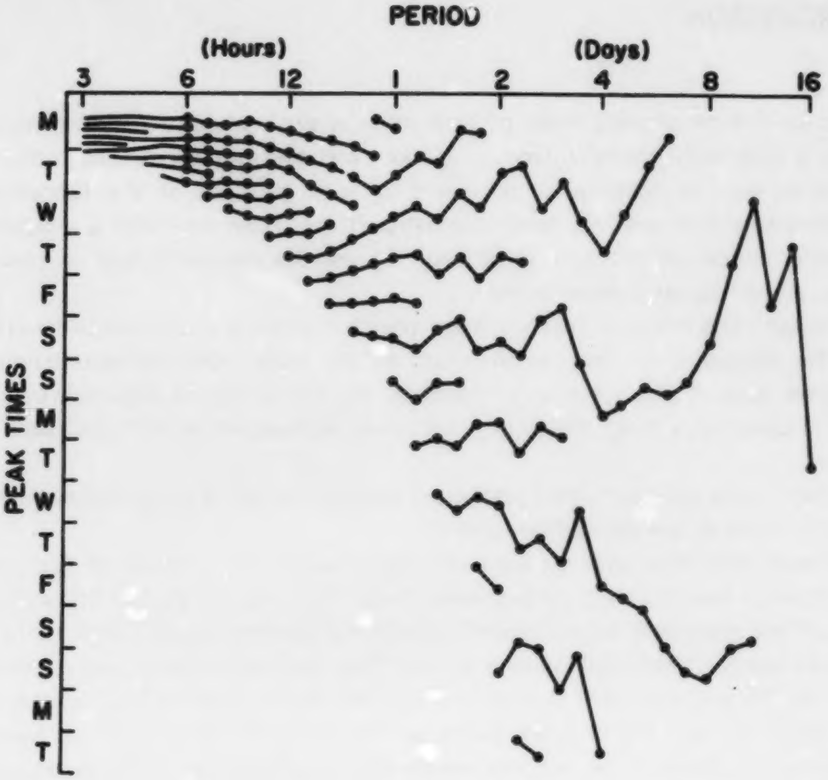


Figure 11-7. Fourier peak time data for concentrations of sulfur dioxide in Cincinnati for October, 1968.

References

Cholak, J., Schafer, L. J., and Yeager, D., 1968: The air transport of lead compounds present in automobile exhaust gases. *Amer. Industrial Hygiene Assoc. J. 29:* 562-8.

Horlick, G., 1971: Fourier transform approaches to spectroscopy. *Anal. Chem. 43:* 61A-66A, July.

Horlick, G., 1972: Digital data handling of spectra utilizing Fourier transformations. *Anal. Chem. 44:* 943-7.

NAPCA, 1969: *1968 Data Tabulations and Summaries, Cincinnati, Continuous Air Monitoring Projects.* National Air Pollution Control Administration, Raleigh, N. C., Publication No. APTD 69-16.

Saltzman, B. E., 1970: Significance of sampling time in air monitoring. *J. Air Pollution Control Association. 20:* 660-5.

Schnelle, K. B., Jr., and Neeley, R. D., 1972: Transient and frequency response of air monitors. *J. Air Pollution Control Association. 22:* 551-5.

DISCUSSION

Marcus: I think this approach of time series analysis of pollutant concentration data is absolutely essential. One thing about focusing on periodicities, is that it is another way of looking at or trying to look at some of the fundamental mechanisms that produce these diurnal patterns. Have you tried a logarithmic transformation of the concentrations? To use the concentrations as they are gives you a Fourier analysis of the . . .

Saltzman: The virtue of using a linear transformation is that is simplifies adding all the components. One can then add all the amplitudes and reconstruct the original data. I don't see any advantage to converting to logarithms. It just complicates everything. You can get an exact representation with the linear data input.

Marcus: Have you done any statistical analysis to test the significance or the reality of the existence of these peaks?

Saltzman: No. What you see are preliminary results. As a matter of fact I want to mention how the calculations were made. You may laugh, but this was done on a Wang computer with a plotter output. To run one month's data required 7 million steps. It took 100 minutes on the Wang to execute, but it was convenient for me. We are now putting it on the IBM/65 which is about 500 times as fast. We hope to get this program going by about March. What you see now are preliminary results. I can say that we are getting spectra and that they do persist for data time periods as long as 3 months.

Marcus: One advantage perhaps of going into a transformation of the data would be to reduce the concentration observations to a somewhat more nearly Gaussian-distributed form and then you could . . .

Saltzman: This procedure has nothing to do with statistics. This is an analytical representation. I am not talking about probabilities here. This is an exact representation. If you use linear terms, you can add and subtract everything.

Marcus: We have an exact representation of intrinsically noisy data and perhaps just transforming that, and trying to again extract a signal, we can get rid of some of the uncertainties that are built into the observations at the beginning.

Benarie: I am very impressed with this spectral representation of air pollution data. It is a great idea. Some amplitudes can be caused by meteorological factors or by human activity. For instance, the 24-hour peak amplitude is almost certainly produced by the early morning inversions. It appears in the Fourier spectrum of every pollutant. The human activity has a weekly cycle, so it can be easily recognized in the amplitudes, even if the data are not for a long period. Secondly, such analysis gives us immediately the answer, following Shannon's communication theory on what should be the sampling frequency of the apparatus. It should be 2N if the highest frequency is N. So we don't have to ask any more if the best sampling time is 5 minutes or 30 minutes or 24 hours. We get the answer out of these spectra.

For the lead, you took 2- and 3-day data. As the concentrations are related to sampling time, an artificial sampling component which is not real has been introduced. All sampling times should be either uniform or random, but not 2 and 3.

Saltzman: In this particular case each data item was weighted for the length of its sampling period for calculations. Now with regard to the proper sampling time, in my paper published in the October, 1970 issue of the Journal of the Air Pollution Control Association, the viewpoint was that if we are interested in the effects on the body, then we are only interested in the frequencies that the body can see. For contaminants with a long biological half-life such as lead (which could be several years), high frequency fluctuations are attenuated and determining the high frequency components is a waste of money. So we would only sample to determine the significant frequencies remaining after attenuation by the biological window through which the body views the data if this is the purpose of sampling.

12. THE PREDICTION OF HIGH CONCENTRATIONS OF SULFUR DIOXIDE IN LONDON AND MANCHESTER AIR

F. BARRY SMITH AND G. H. JEFFREY

Meteorological Office, Bracknell, United Kingdom

Introduction

High concentrations of sulfur dioxide (SO_2) in the atmosphere can cause considerable upset to people with bronchial troubles, particularly if the concentrations are maintained over a period of days. One of the most unpleasant results of the famous London smog of 1952 was the very high mortality rate caused by bronchitis and other related ailments during the subsequent week; overall it was estimated the smog caused between 3,500 and 4,000 deaths (see "Air Pollution and Health", 1970). Hospital places were also in tremendous demand by less seriously affected sufferers.

Quoting from the same source, absenteeism tends to rise rapidly among London factory and office workers whenever the daily average SO_2 concentration exceeds 250 $\mu g/m^3$ (not a particularly high value in London) and, in Salford, absenteeism is twice that of the daily average amongst all workers when the concentration reaches 1000 $\mu g/m^3$ (a rather more exceptional level).

Even the most cursory investigation of weather conditions on days of high SO_2 concentration reveals that cold and relatively calm days in winter are frequently the most dangerous. The Meteorological Office in London was therefore asked over 8 years ago to provide a forecasting service of those meteorological conditions which were likely to lead to significantly high concentrations and a subsequent demand on hospital beds. The criterion chosen was for a concentration of at least 1000 $\mu g/m^3$. During the 1952 smog the maximum daily average SO_2 concentration over 10 sampling stations was approximately 2000 $\mu g/m^3$; however the effects of the Clean Air Act are such that this is about double the greatest SO_2 concentration experienced from 1968 to 1970 inclusive when averaged over 4 stations, with a typical mean somewhat above the Inner London average. In the original scheme developed to meet this demand, the meteorological conditions which were expected to lead to critical concentrations were as follows:

(a) an expectation of less than 2/8th of cloud, or of sky obscured by fog at 18Z, 00Z and 06Z.

(b) an expectation of a mean of surface wind speeds at 18Z, 00Z and 06Z of less than 3 knots, the actual speed at each of these hours being less than 5 knots,

(c) an instability index $S = (2T_x - 3T_n - 12) \geqslant 0$, where T_x is the highest temperature expected at midnight at Crawley at any level up to and including 900 mb (but excluding the surface) and T_n is the forecast minimum temperature at Heathrow (temperatures are in °C). If (a), (b) and (c) are satisfied, a forecast of high pollution is issued, but if in (c) we only have $-3 \leqslant S < 0$ then a more cautious forecast is made.

The London Weather Centre, which is responsible for making the forecasts, has felt that a re-appraisal of the scheme is called for, partially because the scheme did not appear to be highly successful and partially because the Clean Air Act has reduced the overall SO_2 low-level emission rates.

With growing concern all over the world over the state of urban environments, many alternative forecasting schemes have been developed, and several of these are reported in the literature. These generally fall into one of three groups.

(a) Numerical models. Whenever source distributions are reasonably well known both in time and space, the equations of diffusion may be used to calculate spatial distributions of pollution, provided the wind and turbulence characteristics can be adequately prescribed and predicted. Such calculations require considerable computer facilities, and can only be meaningful on a scale that is large compared with the typical distance separating those sources which are not individually represented but are merged into area sources.

(b) Physical models. Detailed models of urban areas have been created in large wind tunnels and the dispersion of pollution emitted in life-like manner from one or more sources studied. The advantage of this system is that the proposed addition of a new major source into an urban environment can be studied fairly realistically, even when the local topography is quite complex. Perhaps their chief disadvantage lies in the difficulty in simulating the wide range of meteorological parameters that affect dispersion: low-level inversions, fogs, solar radiation, wind direction and so on. Their use is therefore more in the urban planning field than in routine day-by-day predictive work.

(c) Empirical models. The scheme outlined above is one such model. The physics of the whole dispersion process only enters in at a comparatively low level, but the scheme does have the advantage that it is based on real data taken in real situations. Considering the very considerable complexity of the problem in an urban environment, the empirical approach may be the only really practical one on a day-by-day basis whenever a sufficient body of data is available for post-facto analysis (say at least 2 years of measurements of SO_2 and the weather). Since such measurements are readily available in London, our revised scheme described in later sections is also of this type.

The Measurements of Sulfur Dioxide in London

Inner (central) London as defined by Weatherley and Gooriah (1970) comprises an area 30 km by 20 km encompassing Hendon in the NW, Dagenham in the NE, Sidcup in the SE and Wimbledon in the SW. Within this area the National Survey sampling network has nearly one hundred sites in operation (the exact number varies between 90 and 100 from year to year.) The area contains industry, scattered mainly around the River Thames and along the Lee Valley, as well as housing and commerce regions with substantial fuel consumption. Parks and comparatively low density housing areas (less than 5000 people per square km) are also present, so that the source distribution and the actual concentration distribution are far from simple (see Figs. 1, 2 and 8). Inspection of the Figures shows that the correspondence between source and concentration, as represented, is not particularly strong on a scale of 1 or 2 kilometers, but is much better on a scale of 5 to 10 kilometers. This perhaps indicates that individual sites may often be significantly influenced by one or two fairly dominant local sources, and only when the concentrations are averaged over, say, four or more sites do they begin to have an obvious meaning in relation to broad area source-values. Figure 1 shows population density and the main industrial areas and comes from Weatherley and Gooriah. Figure 2 shows values of the mean SO_2 winter-values derived from the ten yearly values for each Inner London station in which the smoothed overall trend over the period is linearly extrapolated one year to 1969-70. The mean for all stations is 231 $\mu g/m^3$; however the area-density of stations is not uniform and if isopleths of mean concentration are drawn (ignoring all the possible pitfalls in doing this) the mean concentration determined on an area basis is approximately 213 $\mu g/m^3$. The overall pattern appears to change little from year-to-year but on a shorter time scale significant changes from day-to-day probably occur due to changes in source strength and wind direction. If Figure 2 is representative, concentrations within Inner London vary from at least half, to twice the area mean on any occasion. The highest values are in Westminster, where, since industrial undertakings are few, road traffic and office-block central heating systems may be the most significant polluters of the urban environment.

Figure 3 shows two concentration-direction roses, one for Kensington (site 4), the other for Deptford (site 3). The radius in any direction represents the smoothed mean concentration, relative to the mean for all conditions, when the wind is coming from that direction. An almost 3 to 1 variation in mean concentration with wind direction is implied at both stations, and this appears to be fairly typical.

Meteorological Parameters for London

The analysis of SO_2 concentrations at a rural site which preceded the present London analysis, revealed that day-to-day values depended significantly upon the following parameters:

(a) wind direction. Effective source strengths may vary sufficiently with direction as exampled in the last section.

(b) temperature. Source strength in the UK tends to be greater at lower temperatures. Temperature is also correlated with other meteorological factors that influence the dispersion of the SO_2.

(c) wind speed. Wind speed affects the stability of the atmosphere and hence the vertical dispersion of SO_2. For a specified emission rate of SO_2 the concentration immediately downwind of the source tends to be inversely proportional to the wind speed. It is probable that when ventilation by the exterior wind significantly affects offices and homes, the production of SO_2 increases, following the increase in compensatory heating. Some of these trends are clearly in opposing directions and, at the rural site investigated, were almost self-canceling. In London itself wind speed appears to remain important, particularly at light winds when accumulation of SO_2 within the same mass of air leads to the highest concentrations recorded.

(d) mixing depth or stability. Dispersion through the vertical of SO_2 depends on the intensity of vertical turbulence. Quite frequently a layer near the ground which is well mixed by turbulence is "capped" by a thermal inversion which inhibits further spreading of the pollutant to greater heights. The pollutant is thus trapped, and concentrations tend to a value inversely proportional to the height of the inversion. At places well away from the major source of pollution, the mixing depth is one of the most important parameters, since the approach to uniform mixing below the inversion has time to take place. Within London itself where the typical distance between source and receptor is much less, the mixing depth ceases to have this importance, except when it is very small. (See (d) below.)

The post-facto meteorological data have been obtained from Kew records in Parts I and II, and from London Airport in Part III of the forecasting scheme.

After consideration and experiment it seemed that the most relevant parameters could be defined as follows:

(a) wind direction. 10 meter wind directions, using the tabulated mean over the preceding hour, averaged over 12 hours centered at 15Z during the day when the concentration sample is started. (National Survey 1-day samples start in the morning at an assumed time between 09Z and 10Z and finish 24 hours later). If the wind direction varied by more than 60° during the period, the direction is described as "Variable" and treated as a separate category. Further if there are at least 3 hours of calm (wind speed effectively zero) during the period direction is described as "Calm" and treated as a further separate category.

(b) temperature. In Parts I and II of the forecasting scheme the minimum hourly temperature, during the period 10Z to 24Z on the day when the sample is started, is used. The reasons for this choice are:

(i) temperatures after midnight are not expected to be very relevant since emission rates are then normally quite low.

(ii) the minimum temperature is likely to be well-correlated with the overall coldness of the late afternoon and evening, and hence the domestic heating output.

In Part III of the forecasting scheme the minimum temperature for the whole 24-hour period is used.

(c) wind speed. Two wind speed parameters are extracted. The first is the number of hours when the hourly-mean wind speed (10-meter value) is 2 knots or less (Parts I and II) or less than 5 knots (Part III). For simplicity we call this the number of hours of calm. The second parameter is the mean wind speed for the full day on which the sample is started. Locally a mean speed over the precise period of the sample should have been taken but the sidereal-day mean was already tabulated and thus saved quite an amount of laborious computation at the expense of some accuracy.

(d) the mean reciprocal mixing depth (MRMD). The London analysis indicates that only in situations with low mixing depths, did the MRMD become significant as a predictor. During the winter months either of the following criteria almost always are necessary and sufficient for a significant MRMD:

(i) Surface inversion sets in before 18Z, and during the day cloud height at or below 500m, or

(ii) Surface inversion sets in between 18 and 21Z, and during the day inversion or cloud height at or below 300m.

The rules for surface inversions during the winter are:

(i) At 18Z, assume a surface inversion unless wind speed > 8 kts or cloud amount > 5/8ths.

(ii) At 21Z and 24Z, assume a surface inversion unless wind speed > 8 kts or cloud amount 8/8ths.

The SO_2-Concentration Data

Ideally all sampling stations in the Inner London area should have been used in the analysis. However certain factors weighed against this. For various reasons not all stations maintain a regular day-by-day sampling routine. Further it was decided in this exploratory analysis to limit the amount of data to that which could be handled and analyzed fairly easily using a desk electronic computer, the Olivetti Programma 101.

Consequently 4 stations with a good record of completeness were selected, and permission to use their data was kindly granted by the Councils concerned. These stations are:

Kensington, Site 4
City of London, Site 17
Hackney, Site 2
Deptford, Site 3.

Mean concentrations for a particular day were evaluated whenever either 3 or 4 of the stations gave readings. In the former case the mean was given the appropriate weighting to balance the omission of one of the readings:

Expected mean concentration when C_4 is missing

$$= 1/4\,(C_1+C_2+C_3)\left(\frac{m_1+m_2+m_3+m_4}{m_1+m_2+m_3}\right)$$

where C_1, C_2 and C_3 are the day's readings at the 3 given sites; m_1, m_2, m_3 and m_4 are the long-term mean concentrations. For the 2 winter periods that are studied in detail in this analysis (the winter of 1968-69 and that of 1969-70) they take the following values:

m (Kensington) = 364 $\mu g/m^3$
m (City of London) = 415 $\mu g/m^3$
m (Hackney) = 376 $\mu g/m^3$
m (Deptford) = 253 $\mu g/m^3$

Winter covers the months from October to March inclusive.

No readings were taken on Saturdays or Sundays, and Monday's readings represent combined values for the 3 weekend days. Three days out of 7 are therefore not available for the present analysis. Public holidays are also sometimes missed. In all, 194 days had 3 or 4 readings at the sites and this is in fact a very high proportion of the total possible number of days.

The mean concentrations are higher by some 11% than the winter averages for the 2 years given in the annual Warren Spring Laboratory Reports, "Investigation of Air Pollution", largely it seems because the omitted weekend concentrations are on average lower than the midweek concentrations. The two-year winter averages, on the other hand, are satisfactorily close to the five-year winter averages.

Concentrations at the 4 sites are not of course perfectly correlated on a day-to-day basis, partially because changes of wind direction change the source distributions which affect each sampler, and partially because of normal variations in source output from each and every source. The correlation coefficients between the concentrations at the sites vary from about 0.44 to 0.68. Now if we may assume that $C/\bar{C}$ (where $\bar{C}$ is the time-mean concentration at one site, and C is a 1-day concentration at the same site) has a statistical day-by-day distribution which is virtually the same irrespective of site, then

$$\sigma^2 = (1-r)\,s^2$$

where σ = the standard deviation of the "random" component of the concentration C, which is uncorrelated from site to site

r = the site-to-site correlation coefficient

s = the standard deviation of the concentration values at any site.

Typically then, $r \approx 0.56$ and $s \approx 180\ \mu g/m^3$. Roughly, we deduce that $\sigma \approx 100\ \mu g/m^3$. This implies that the concentration at any site on any day cannot be specified, even when the Inner London mean concentration is known, to within an error e which has a standard deviation $\sigma \approx 100\ \mu g/m^3$.

Averaging the concentrations over 4 sites reduces this error by $4^{1/2}$, i.e., the standard error is now $50\ \mu g/m^3$. Averaging over all the 94 sites would reduce the standard error further to about $10\ \mu g/m^3$. The random error for the 4 sites must be one of the reasons for the failures, albeit a relatively small number of failures, in the forecasting scheme described later.

Figure 4 shows the histogram of 290 mean winter concentrations for 5 years for all the Inner London stations, when means could be evaluated, taken from the Warren Spring Laboratory Annual Reports (loc. cit.). The histogram conforms quite closely to a lognormal distribution with a median of $235\ \mu g/m^3$ as shown on Figure 4. The mean (including weekends) of the 4 stations is $310\ \mu g/m^3$, and thus some 20% of Inner London may be expected to experience concentrations greater than the average of the 4 stations on a winter basis, and with less certainty on a daily basis. If the lognormal hypothesis is correct, some 0.2% (i.e., 1.2 sq km) of the total area may experience twice the 4-station average.

To separate more clearly the spatial and temporal distributions of concentration, Figures 5 and 6 show the cumulative frequency curves for the concentrations meaned over the five winters 1965-70 for all the Inner London sites for which values could be obtained, and for the daily concentration values, meaned over the four sites, for the two winter periods under survey, 1968-69 and 1969-70. Both curves show a close approximation to lognormal distributions.

To support the hypothesis that the lognormal distribution is a satisfactory fit, at least over the inner 90% of the distribution, the following test may be applied:

If C_m = the median concentration of the distribution

$\bar{C}$ = the mean concentration of the distribution

σ = the standard deviation of lnC

s = the standard deviation of C

then for a lognormal distribution:

$$\bar{C} = C_m \exp(1/2\,\sigma^2) \qquad (1)$$

and

$$s^2 = (\bar{C}^2 - C_m^2)\,\bar{C}^2/C_m^2 \tag{2}$$

C_m is calculated by forming the geometric mean of all the concentration values in the sample and is a theoretically better estimate of the parent population median concentration than is $\bar{C}$, the arithmetic mean, of the parent population mean concentration. Similarly σ is more reliable than s.

Applied to the 73 data values involved in Figure 5:

FROM THE DATA	CALC. FROM EQ. (1) and (2)
$C_m = 227.1$ $\bar{C} = 238.6$	$\bar{C} = 238.0$
$\sigma = 0.313$ $s = 76.8$	$s = 74.6$

The median evaluated by its fundamental definition, namely by the value which equally divides the data points (50% having a higher concentration and 50% a lower) is $C_m = 231$ μg/m^3. However this is a less accurate method of estimating the parent population C_m from a sample on the assumption of a lognormal distribution.

The close agreement between the calculated and derived values of $\bar{C}$ and s strongly support the lognormal hypothesis. The advantage of this hypothesis is that it enables us to estimate the likely area in Inner London in which the concentration of SO_2 may exceed some defined critical level at any time. However the hypothesis must remain of doubtful validity "out on its tails", i.e., when the area becomes smaller than about 10 sq km, and too much reliance should not be placed on forecasts in these circumstances without a much more detailed investigation than is given here.

One final point concerning these statistics may be made. The geographical distributions of

(a) the mean concentrations for the winters under analysis, and

(b) the number of days with concentrations exceeding 500 μg/m^3 shown in Figures 7 and 8, are very similar. The following approximate correspondence apply:

Number of days with $\bar{C} \geqslant 500$ μg/m^3 over two winters	Mean concentration for the same two winters (μg/m^3)
0	150
10	200
25	300
50	360
100	400

These relations should be roughly consistent with the lognormal time distributions of concentration.

The Data for Manchester

Sulfur dioxide data were obtained for the same 2 winters as for London from 7 regular sampling stations in the National Survey Network in Manchester: Numbers 11, 13, 15, 16, 17, 18 and 19.

Meteorological data came from Manchester Airport (Ringway) some 9 miles to the south of the city.

Both sets of data were treated in the same way as in the analysis of the London data and thus no further explanations will be given.

The Variation of Concentration with Meteorological Parameters

The previous section headed "Meteorological parameters for London" described the meteorological parameters that appeared to be significant.

Table I gives the variation of mean concentration, averaged over the 4 sites, with wind direction.

TABLE I. Variation of Mean Concentration With Wind Direction

Wind Direction	Mean concentration, μg/m^3	Wind Direction	Mean concentration, μg/m^3
001-030	243	181-210	235
031-060	271	211-240	204
061-090	351	241-270	223
091-120	395	271-300	232
121-150	302	301-330	323
151-180	268	331-360	279
Variable	307	Calm	306

Table II sets out in detail all the basic data, some of which has already been defined in section "Meteorological parameters for London." The column headed MRMD gives the mean reciprocal mixing depth described as H when it is significantly important. The penultimate column represents the results of the objective post-facto forecasting scheme (Part I).

The forecast scheme was developed empirically by considering the concentration values and the appropriate meteorological parameters for the first winter 1968-69. When applied to the second winter 1969-70, the scheme proved to be equally successful without any further modification or elaboration of the rules. The rules may be stated quite simply as follows.

The Forecasting Scheme: Part I

(a) A concentration averaged over the usual 24-hour period at the 4 stations: Kensington 4, City of London 17, Hackney 2, Deptford 3, will exceed 400 μg/m^3 (or in the case of those wind directions which on average have low

SO_2 concentrations, a normalized concentration exceeding 1.5), whenever at least one of the following conditions is fulfilled:

(i) the number of hours with mean wind speed less or equal to 2 knots greater or equal to 8 hours (see column 5, Table II)

(ii) minimum temperature (col 7) $\leqslant 0°C$, and at least 1 hour light winds (col 5)

(iii) the MRMD (col 8) = H and minimum temperature (col 7) $\leqslant 6°C$, and mean wind (col 6) $\leqslant$ 10 knots

(iv) $(3\,t_{calm} - 2T_{min})$ is between 0 and 25 if c (previous day) > 600 $\mu g/m^3$, or between 10 and 35 if C (previous day) $> 400\ \mu g/m^3$

(b) The concentration defined in (a) above will exceed 600 $\mu g/m^3$ whenever

(i) the minimum temperature (col 7) is less than 5°C; and light winds (col 5) for 19 or more hours

(ii) $(3t_{calm} - 2T_{min})$ exceeds 25 if C (previous day) $> 600\ \mu g/m^3$, or exceeds 35 if C (previous day) $> 400\ \mu g/m^3$

The results displayed in Table II may be summarized in the following tables:

(A) Contingency Table for success in forecasting A*

(i.e., London: either C $\geqslant 400\ \mu g/m^3$, or normalised C $\geqslant 1.5$; Manchester C $\geqslant$ 270 $\mu g/m^3$)

	high concentration forecast A		lower concentration forecast not A		Total	
	London	M/C	London	M/C	London	M/C
forecasting success =	55 28%	49 21%	106 54%	144 61%	161 82%	193 82%
forecasting failure x	15 8%	20 9%	17 9%	22 9%	32 18%	42 18%
Total London	70 36%		123 64%		193	
M/C	69 30%		166 70%			235

Exactly equivalent information is included for the Manchester data, without giving the basic data equivalent to Table II.

In both cases when a forecast of high pollution is made, a success rate of about 80% is achieved.

Some important points must be made:

(a) The London threshold values 400 and 600 $\mu g/m^3$ are not universal values. They are only meaningful in so far as the source distribution and output remains basically unaltered. While it is virtually impossible over a short period of time to identify any such change, it is recommended that the two values be

(B) Contingency Table for success in forecasting B*
(i.e., $C \geqslant 600\ \mu g/m^3$ for London; $C \geqslant 450\ \mu g/m^3$ for Manchester)

	very high concentration forecast B		lower concentration forecast not B		Total	
	London	M/C	London	M/C	London	M/C
forecasting success =	11 6%	11 5%	182 93.5%	212 90%	192 99.5%	223 95%
forecasting failure x	0 0	8 3%	1 0.5%	4 2%	1 0.5%	12 5%
Total	11 6%	19 8%	182 94%	216 92%	193	235

*Percentages in general have been rounded to the nearest whole number.

suitably modified if necessary once every 5 years in the light of the overall changes in mean winter concentration at the 4 sites over the preceding 5 years.

(b) No attempt has been made in this analysis to relate the concentration values to the effect on people's health and the likely demands on the facilities at the two hospitals concerned. This is largely a medical problem and lies outside our capabilities.

(c) In the previous scheme a forecast had to be made before 1600Z of the chances of high pollution during the evening and night that followed. We have moved to a different problem, partially because our basic concentration data are daily mean values (rather than hour-by-hour values) and also because we feel that the problem is not solely a night-time problem. At night many people, and particularly bronchitic sufferers, are likely to be in the shelter of their own homes where they can to some extent regulate the condition of the air they breathe, whereas during the day they are more likely to be out and about, being affected by atmospheric concentrations of SO_2 which are not necessarily a great deal less than the evening concentration. Our aim has therefore been to forecast the mean concentration for the whole 24-hour day. The forecast of the meteorological conditions is therefore longer-range and to that extent more liable to error.

Forecast	The percentage of forecasts made by the scheme that were correct		The percentage of actual cases correctly forecast by the scheme	
	London	Manchester	London	Manchester
$C > A$	79	71	76	69
$C < A$	86	87	88	89
$C > B$	100	58	92	73
$C < B$	99	99	100	98

For London: $A = 400 \mu g/m^3$, $B = 600 \mu g/m^3$
For Manchester: $A = 270 \mu g/m^3$, $B = 450 \mu g/m^3$

Ideally, then, a forecast should be made in the early morning, before 1000Z, of the meteorological conditions for the next 24 hours and hence the likely mean concentration. Since many of the criteria in the forecasting scheme relate to evening conditions (and hence are not altogether different in intent from the previous shceme), some revision of the concentration forecast could be made as late as 1600Z whenever this seemed called for.

(d) The results set out above refer to a post-facto application of the scheme using meteorological data as it actually occurred. In day-by-day application of the scheme in the future these data will have to be forecast and this is bound to introduce further significant errors.

Some of the parameters, such as the minimum temperature and the cloud amounts, are already estimated on a routine basis for other purposes. The criterion of the number of hours when the mean wind falls to 2 knots or below is probably the hardest to estimate with any certainty, and for this reason Part III explores the effect of a relaxation of this condition.

A scheme designed to forecast actual concentration values: Part II

This part is concerned with predicting actual concentration, as distinct from forecasting whether or not certain threshold values are exceeded. The variation of concentration with the same meteorological parameters that were successfully used in the "threshold" method was studied for the 2 winter periods for the London data. The following fairly simple formula yielded reasonably satisfactory estimates of the average daily concentrations at the 4 sites:

$$C_{est} = \left(1 + \frac{\delta_d}{3}\right)\left(1 + \frac{\delta_m}{6}\right)\left(1 - \frac{T}{28} + \frac{t}{20}\right)\left(a\bar{C} + bC_p\right)$$

where $\bar{C}$ = long term mean concentration

C_{est} = estimated concentration (24-hour average)

C_p = concentration for the previous 24 hours

T = minimum temperature (°C) expected up to midnight

t = number of hours of mean wind less than 3 knots during the 24 hours

$a = \frac{2}{3}$, $b = \frac{2}{9}$ if the mean wind for the day exceeds 6 knots

$a = \frac{3}{7}$, $b = \frac{8}{21}$ otherwise

δ_d = 1, if the mean wind comes from the "dirty" sector, 060° to 120°
0, otherwise

δ_m = 1, if the mixing depth is low (as defined at the end of section on London)
0, otherwise

Although the formula has been verified only for the 4 sites, it is probably equally applicable to any group of sites in Inner London, and can be applied to other cities provided the appropriate value of C, the mean concentration, is inserted. This has been done for the Manchester data.

The formula can either be expressed graphically as a nomogram (see Figure 9) or it can be programmed for a desk electronic calculator.

The root-mean-square errors have been evaluated for the 4 sites over the 2 winters. The significance of the errors have to be assessed in relation to the inherent "error" due to local quasi-random variations in concentration at the 4 sites, which can be estimated from the inter-site correlation studies described in the section about SO_2 concentration data. The inherent error in the 4-site average concentration was shown to be about 50 $\mu g/m^3$.

The root-mean-square error in the formula-estimates is only 66 $\mu g/m^3$ (little more than the inherent error) if the actual value of C_p, the previous day's concentration, is known and used, but rises to nearly 80 $\mu g/m^3$ when C_p is only known by the application of the formula using the actual meteorological data at the end of the previous 24-hour day (see Figure 10). This is still a satisfactorily small margin of error when compared with the inherent error, and is certainly considerably less than that obtained using a persistence forecast (142 $\mu g/m^3$) i.e. by using yesterday's concentration as an estimate for today's. If E_t is the total error, E_i is the inherent error, and E_s is the basic error of the scheme, then

$$E_t^2 = E_i^2 + E_s^2$$

For the whole of Inner London (nearly 100 sites), E_i will fall from 50 $\mu g/m^3$ to about 10 $\mu g/m^3$. The expected value of E_t would then be

$$E_t^2 = (80)^2 - (50)^2 + (10)^2$$

i.e. a little over 60 $\mu g/m^3$.

The formula displays the relative importance of the basic parameters. It is clear that an error of 4°C in T, the minimum temperature, would introduce an error in C_{est} of about only 50 $\mu g/m^3$. A similar error would follow from an error of 3 hours in t, the hours of light winds, or of about 150 $\mu g/m^3$ in C_p. The method does not therefore demand impossible precision in evaluating the basic meteorological parameters. Nevertheless if it has to be used at an operational office, such as the London Weather Centre, evaluation of these parameters may take rather more time than the fully occupied staff may wish to spend. The next section (Part III) describes a simplified scheme designed to overcome this difficulty.

The simplified forecasting scheme: Part III

In order to help the operational forecaster by minimizing the analysis required in making a pollution forecast, an investigation has been made into the effect on the accuracy of the scheme when:

(a) London Airport meteorological data is used instead of Kew data for London. Unlike Kew, London Airport data is received by London Weather Centre on an hourly basis.

(b) The "hours of calm" criterion is relaxed to include all hours when the mean wind falls below 5 knots. This criterion should be much easier to forecast.

(c) The minimum temperature up to midnight is replaced by the minimum temperature over the 24 hour period of the forecast. The forecaster will already have this temperature estimate for other reasons.

(d) The effects of the daily mean wind speed and direction are ignored.

The empirical formula for forecasting concentration in terms of the revised parameters is

$$C_{est} = 0.085\left(1 + \frac{\delta_m}{6}\right)\left(1 - \frac{T - t}{28}\right)\left(5\bar{C} + 4C_p\right) + 0.15\bar{C}$$

where δ_m = 1 if MRMD = H, and is 0 otherwise
T = minimum temperature 09Z to 09Z
t = hours when mean wind falls below 5 kts
$\bar{C}$ = mean concentration
C_p = yesterday's concentration

The results using this scheme are summarized in Table III. As expected the errors are somewhat bigger than in Part II, but not appreciably so. It seems that this very simple scheme still gives a very satisfactory means of forecasting pollution.

For many cities, including Manchester, the number of days with a low mixing depth (MRMD = H) is very small (only a few days per year), and experience may show that the factor $(1 + \delta_m/6)$ can be then fairly safely ignored.

TABLE II. Basic Forecasting Data

Date	C[a]	dd[b]	$C/\bar{c}$[c]	t_{calm}[d]	$\bar{v}$[e]	T_{min}[f]	MRMD[g]	F/C[h]	Estimated C[i]
1968									
Oct. 1	127	260	0.57	0	13	14	H	=	–
2	104	240	0.51	0	13	15	–	=	121
3	121	240	0.59	0	14	15	–	=	119
4	198	260	0.89	4	9	16	–	=	164
10	172	220	0.84	1	8	14	–	=	–
11	123	240	0.60	0	13	12	–	=	155
15	182	260	0.82	2	8	10	–	=	–
16	174	230	0.85	0	12	11	–	=	166
17	161	260	0.72	0	9	9	–	=	185
18	314	250	1.41	13	7	8	–	Ax	367
22	493	140	1.63	15	5	10	–	A=	–
23	548	360	1.96	18	3	12	–	A=	497
24	480	050	1.77	8	6	11	–	A=	477
25	331	070	0.94	0	10	12	–	=	259
29	253	190	1.08	0	12	12	–	=	–
30	197	230	0.96	7	12	11	–	=	277
31	204	180	0.76	7	8	14	–	=	236
Nov. 1	202	200	0.86	2	12	13	–	=	177
5	418	030	1.72	12	11	2	–	A=	–
6	378	070	1.08	2	9	7	H	=	431
7	311	070	0.89	2	14	7	–	=	360
8	283	060	1.04	2	11	6	–	=	357
12	546	150	1.81	11	6	4	–	A=	–
13	506	110	1.28	0	10	4	–	x	405

[a] Corrected mean SO_2 ($\mu g/m^3$)

[b] Wind direction

[c] Concentration normalized by mean conc. for wind direction

[d] Hours $V \leqslant 2$ kt

[e] Mean wind speed (kt)

[f] Minimum temp (°C)

[g] Mean reciprocal mixing depth (marked H when significantly high)

[h] A: forecast concentration exceeding 400 $\mu g/m^3$, or a normalized concentration exceeding 1.5. The second alternative allows for concentrations below 400 $\mu g/m^3$, which considering the wind directions, are nevertheless high. Normalized concentration is defined as the ratio of the actual concentration to the mean concentration for that particular wind direction.

B: forecast concentration exceeding 600 $\mu g/m^3$.

=: correct forecast

X: incorrect forecast, either a predicted A or B not born out in practice, or no forecast of A or B when there should have been.

[i] Concentration using the empirical formula (Part II) when C available on previous day.

TABLE II (continued). Basic Forecasting Data

Date	C^a	dd^b	$C/\bar{c}^c$	$t_{calm}{}^d$	$\bar{v}^e$	$T_{min}{}^f$	$MRMD^g$	F/C^h	Estimated C^i
14	431	100	1.09	0	14	6	–	x	362
15	394	090	1.12	0	17	5	–	=	344
19	284	340	1.02	3	10	6	–	=	–
20	425	320	1.21	20	5	7	–	A=	519
21	382	150	1.26	5	8	7	H	=	382
22	317	160	1.18	1	9	6	H	Ax	310
26	225	190	0.96	0	12	10	–	=	–
27	250	190	1.06	2	11	12	H	=	222
28	585	120	1.48	17	6	7	H	A=	616
29	529	030	2.18	12	4	8	H	A=	557
Dec. 3	892	080	2.54	12	7	2	H	B=	–
4	443	160	1.65	10	5	4	–	A=	668
5	303	150	1.00	2	6	4	–	=	318
6	481	140	1.59	12	8	1	–	A=	470
10	618	030	2.54	20	5	2	–	B=	–
11	374	050	1.38	1	6	3	–	=	349
12	469	340	1.68	16	7	4	H	A=	612
13	864	060	3.19	13	5	0	H	B=	723
17	419	220	2.05	7	9	0	–	A=	–
18	300	180	1.12	2	10	6	–	=	289
24	278	270	1.25	3	13	6	H	=	–
31	570	330	1.76	12	7	0	–	A=	–
1969									
Jan. 1	677	340	2.43	17	7	2	H	B=	537
2	565	270	2.53	2	8	0	H	A=	493
3	499	330	1.54	14	8	5	–	A=	546
7	500	090	1.42	0	10	1	–	x	–
8	346	150	1.14	0	14	6	–	=	271
9	435	190	1.85	10	12	2	–	A=	443
10	421	120	1.06	0	8	2	H	A=	477
14	321	190	1.36	0	16	8	–	=	–
15	271	190	1.15	0	13	4	–	=	261
16	287	170	1.07	3	10	4	–	=	296
17	248	270	1.11	1	13	3	–	=	280
21	290	150	0.96	0	8	6	–	=	–
22	319	190	1.36	0	12	10	–	=	191
23	351	230	1.72	8	10	11	H	A=	357
24	281	240	1.38	13	5	10	H	Ax	428
28	274	200	1.16	0	12	7	–	=	–
29	353	270	1.58	9	8	3	H	A=	461
30	341	270	1.53	1	8	4	H	A=	330
31	197	240	0.96	0	14	7	–	=	232
Feb. 4	449	340	1.61	4	14	-1	–	A=	–
5	670	280	2.89	17	6	-1	H	B=	733
6	706	280	3.04	16	4	4	H	B=	783
7	329	240	1.61	0	6	2	–	x	362
11	635	320	1.96	5	9	-3	–	B=	–
12	318	V	1.03	0	11	1	–	=	361
13	233	360	0.83	0	14	0	–	=	304

TABLE II (continued). Basic Forecasting Data

Date	C^a	dd^b	$C/\bar{c}^c$	t_{calm}^d	$\bar{v}^e$	T_{min}^f	$MRMD^g$	F/C^h	Estimated C^i
14	503	360	1.80	15	11	-1	H	A=	594
18	379	360	1.36	5	9	-1	–	Ax	–
19	409	050	1.51	0	10	0	–	x	318
20	346	050	1.28	0	21	-1	H	=	392
21	475	180	1.77	1	13	2	–	x	304
25	591	060	2.18	9	10	6	H	A=	–
26	381	040	1.40	0	6	4	–	=	313
27	276	030	1.13	0	9	0	–	=	318
28	289	350	1.03	0	11	2	–	=	274
Mar. 4	387	360	1.39	0	11	3	–	=	–
5	396	040	1.46	0	14	0	–	=	319
6	396	050	1.46	0	14	1	H	=	394
7	581	150	2.14	12	8	-3	–	A=	549
11	552	080	1.57	0	13	4	–	x	–
12	399	V	1.30	0	11	3	–	=	318
13	525	070	1.50	0	10	3	–	x	383
14	216	220	1.06	1	12	8	–	=	312
18	472	070	1.34	0	10	6	H	A=	–
19	498	070	1.42	0	10	6	H	A=	413
20	369	020	1.52	2	6	5	–	x	317
21	372	010	1.53	4	2	5	–	x	297
25	280	360	1.00	0	10	1	–	=	–
26	251	030	1.03	0	10	2	–	=	274
27	171	030	0.70	0	12	1	–	=	279
28	280	050	1.03	4	12	1	–	=	246
Oct. 1	175	280	0.75	4	3	7	–	=	–
2	235	310	0.73	9	4	3	–	Ax	291
3	165	250	0.74	0	7	12	H	=	190
7	195	220	0.95	2	3	15	H	=	–
8	149	210	0.63	2	7	15	–	=	186
9	255	200	1.08	14	5	10	–	Ax	278
10	380	130	1.26	19	1	7	–	Ax	420
14	207	120	0.52	1	5	14	–	=	–
15	177	240	0.97	2	7	9	–	=	217
16	197	190	0.84	1	7	10	–	=	189
17	255	160	0.95	11	6	8	H	Ax	326
21	328	170	1.22	19	1	9	–	Ax	–
22	429	040	1.58	17	1	12	–	A=	391
23	301	250	1.35	12	2	11	H	Ax	441
24	126	250	0.56	0	7	13	–	=	161
28	333	270	1.50	15	2	5	–	A=	–
29	223	200	0.95	4	3	10	–	=	233
30	263	340	0.94	7	7	5	H	Ax	387
31	205	280	0.88	0	6	9	–	=	198
Nov. 4	135	230	0.66	0	17	14	–	=	–
5	224	300	0.96	2	11	5	H	=	283
6	417	270	1.87	13	4	-2	–	A=	405
7	296	160	1.10	4	3	5	–	=	315
11	248	230	1.21	0	9	8	–	=	–

TABLE II (continued). Basic Forecasting Data

Date	C^a	dd^b	$C/\bar{c}^c$	t_{calm}^d	$\bar{v}^e$	T_{min}^f	$MRMD^g$	F/C^h	Estimated C^i
12	163	180	0.61	2	14	12	—	=	194
13	192	220	0.94	0	10	6	—	=	212
14	279	210	1.61	16	4	0	—	A=	402
18	461	290	1.99	9	8	-2	H	A=	—
19	292	210	1.24	0	8	2	—	=	312
20	161	230	0.79	0	11	10	—	=	192
25	320	320	0.99	0	10	1	—	=	—
26	280	310	0.87	0	12	1	—	=	294
27	395	320	1.22	1	9	-1	H	Ax	374
28	287	240	1.41	0	8	5	—	=	264
Dec. 2	529	230	2.59	1	4	2	H	A=	—
3	356	230	1.74	0	7	5	—	x	288
4	218	280	0.94	0	9	4	—	=	268
9	661	330	2.05	6	7	1	—	x	—
10	1161	V	3.78	24	1	3	H	B=	981
11	632	170	2.36	13	1	0	H	B=	1140
12	357	180	1.33	2	4	6	—	=	346
16	304	260	1.36	0	12	3	—	=	—
17	421	340	1.51	7	6	-1	H	A=	486
18	411	060	1.52	0	13	1	—	x	420
19	463	020	1.90	10	9	-2	—	A=	510
23	320	290	1.30	0	12	6	H	=	—
24	318	230	1.56	0	5	4	H	A=	272
30	428	090	1.22	0	12	0	—	x	—
31	265	040	0.98	0	17	0	—	=	328
1970									
Jan. 1	215	030	0.88	0	16	-1	—	=	303
2	387	V	1.26		9	-1	—	=	291
6	453	270	2.03	1	4	-4	H	A=	—
7	556	270	2.49	7	9	-2	H	A=	554
8	659	C	2.15	4	4	-5	H	B=	582
9	438	090	1.25	0	13	-2	—	A=	543
13	284	170	1.06	5	9	5	H	Ax	—
14	249	140	0.82	0	7	8	—	=	212
15	232	140	0.77	2	10	8	—	=	235
16	283	130	0.94	8	4	8	—	=	266
20	178	130	0.59	0	8	7	—	=	—
21	286	160	1.07	7	5	7	H	=	280
22	250	160	0.93	2	9	8	—	=	242
23	295	190	1.25	3	5	3	H	Ax	298
27	423	230	2.07	11	3	4	H	A=	—
29	399	120	1.01	0	6	3	—	=	—
30	257	100	0.65	0	11	5	—	=	353
Feb. 3	173	230	0.85	0	13	7	—	=	—
4	278	250	1.25	0	12	4	—	=	233
5	277	230	1.36	5	9	3	—	=	337
6	317	350	1.14	1	7	2	—	=	314
10	197	250	0.88	0	11	2	—	=	—
11	421	270	1.89	9	8	-4	—	A=	441

TABLE II (continued). Basic Forecasting Data

Date	C^a	dd^b	$C/\bar{c}^c$	$t_{calm}{}^d$	$\bar{v}^e$	$T_{min}{}^f$	$MRMD^g$	F/C^h	Estimated C^i
12	505	V	1.67	3	4	-1	H	A=	429
13	318	030	1.31	2	16	1	—	=	368
17	366	300	1.58	6	8	-3	H	A=	—
18	253	270	1.13	0	13	3	—	=	281
19	243	230	1.19	0	9	4	—	=	248
20	158	220	0.77	0	12	8	—	=	205
24	247	250	1.11	0	8	6	—	=	—
25	269	290	1.16	2	9	3	—	=	286
26	307	330	0.95	5	5	1	—	=	307
27	239	350	0.86	0	9	2	—	=	280
Mar. 3	262	320	0.81	0	10	-1	—	=	—
4	354	300	1.52	0	8	-1	—	x	302
5	340	V	1.11	0	11	0	—	=	312
6	447	320	1.38	0	7	-1	—	x	320
10	683	V	2.22	12	3	-1	H	B=	—
11	495	210	2.11	1	4	1	—	A=	416
12	444	200	1.89	3	8	3	H	A=	418
13	429	070	1.22	0	8	2	—	x	411
17	231	280	0.99	0	7	7	—	=	—
18	240	240	1.18	0	11	10	—	=	183
19	189	270	0.85	0	13	5	—	=	235
20	212	270	0.95	0	11	7	H	=	241
24	282	V	0.92	3	5	6	—	=	—
26	314	010	1.29	8	7	-1	—	Ax	425

TABLE III

Results of applying the Forecasting Schemes to London and Manchester data

	Correlation between actual and forecast concentrations		Standard error, $\mu g/m^3$	
Detailed Scheme: Part II	London	Manchester	London	Manchester
Inherent error from taking only 4 sites:	0.94	—	50	—
All data: actual conc. for previous day known:	0.87	0.80	76	65
As above, excluding 11.12.69:	0.90	—	66	—
Previous day's concentration replaced by mean $\overline{C}$:	0.80	—	95	—
Previous day's conc. replaced by f/c value for that day:	0.85	—	79	—
Persistence f/c: previous day's conc. used as f/c for today:	0.55	—	142	—
Simplified Scheme: Part III				
All data: actual conc. for previous day known:	0.81	0.79	88	66
As above, excluding 11.12.69:	0.84	—	83	—
Standard deviation of all the observations:	—	—	150	106

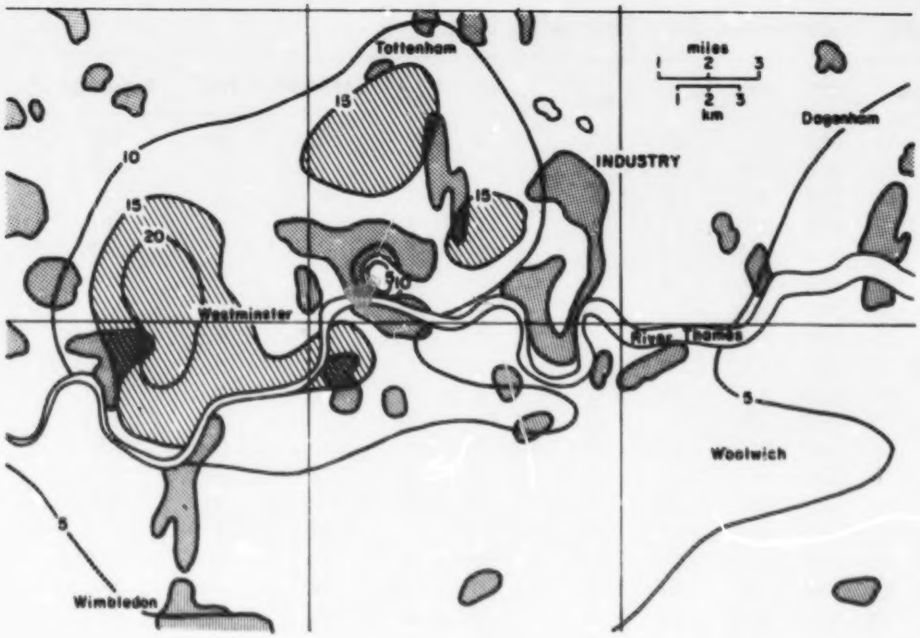

Tottenham
miles
1 2 3
1 2 3
km
Dagenham
INDUSTRY
10
15
15
20
15
10
Westminster
River Thames
5
Woolwich
5
Wimbledon

Figure 12-1. Population density of Inner London in thousands per square kilometer.

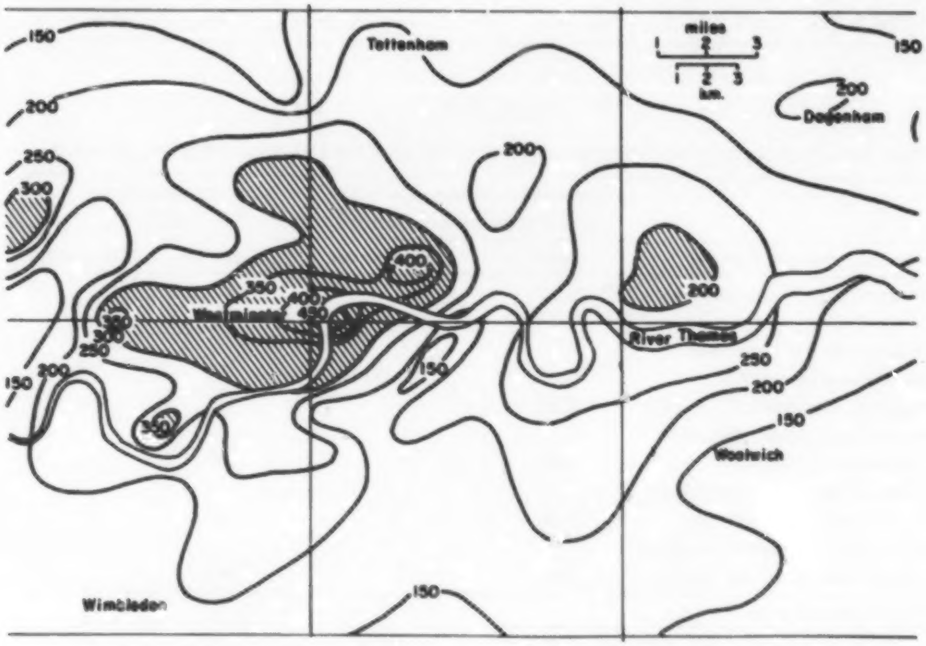

150
Tottenham
miles
1 2 3
1 2 3
km.
150
200
Dagenham
200
250
300
200
400
350
400
Westminster
450
350
300
250
200
150
200
River Thames
150
250
200
150
Woolwich
350
Wimbledon
150

Figure 12-2. Mean winter concentrations for Inner London for 1969-70 based on extrapolation from data of previous ten years. Sulfur dioxide concentrations in $\mu g/m^3$.

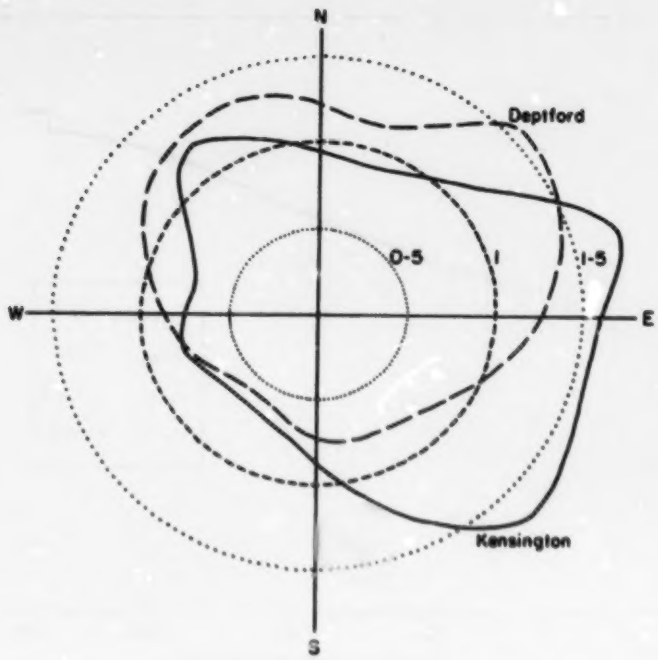


Figure 12-3. Normalized concentration-direction roses for Kensington & Deptford.

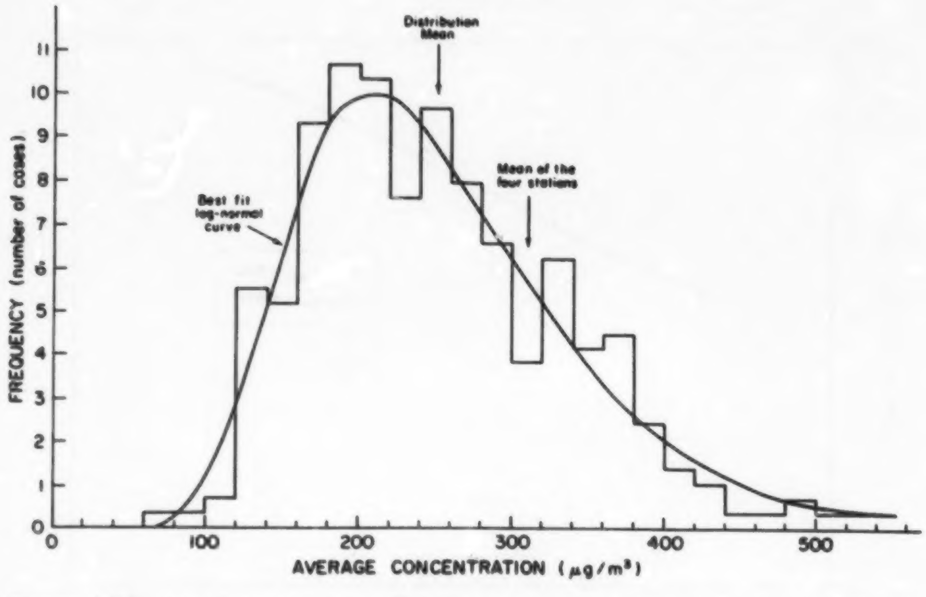


Figure 12-4. Histogram based on 290 values (mean winter concentrations) for all Inner London sites for the winters 1965-66 to 1969-70. Data grouped into 20 $\mu g/m^3$ classes.

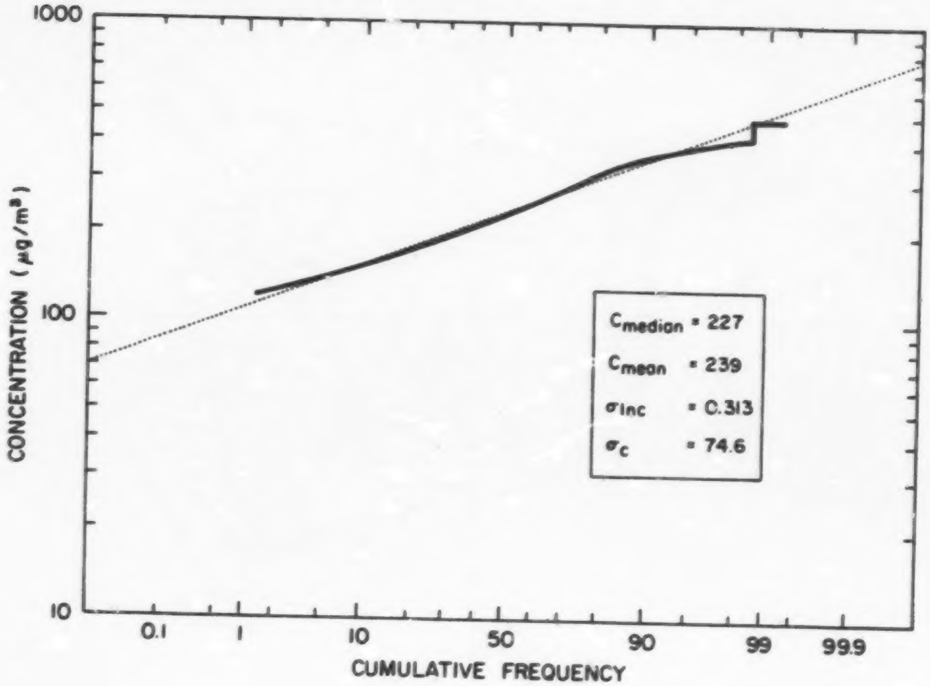


Figure 12-5. Spatial distribution for mean winter 1965-70, all Inner London sites.

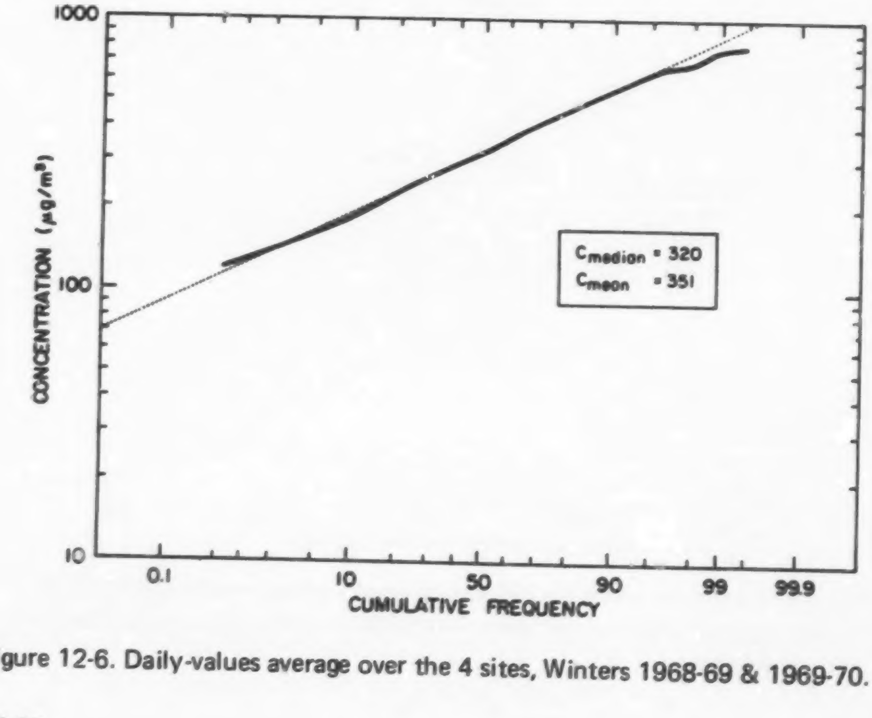


Figure 12-6. Daily-values average over the 4 sites, Winters 1968-69 & 1969-70.

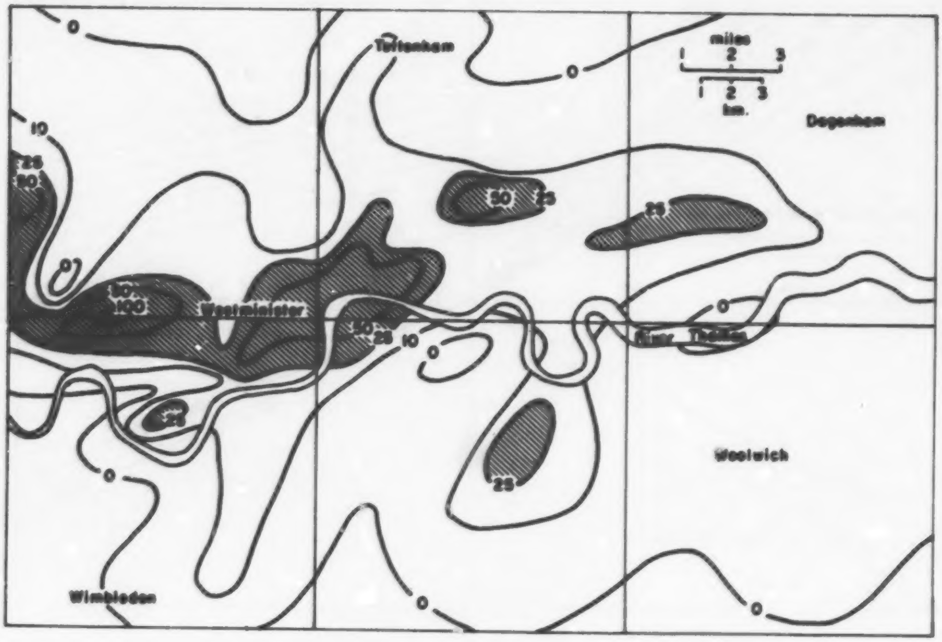

0
Tottenham
miles
1 2 3
1 2 3
km.
Dagenham
10
25
50
50
25
25
0
50
100
Westminster
50
25
10
0
0
River Thames
25
25
Woolwich
0
Wimbledon
0
0

Figure 12-7. Number o´ days when $\bar{C} > 500$ for the winters 1968-69 & 1969-70.

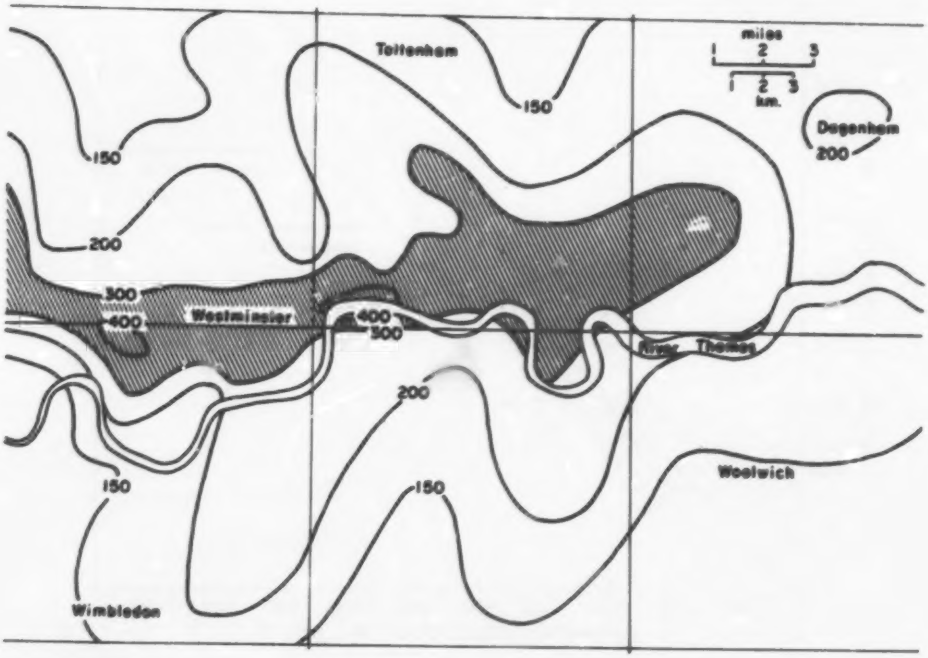
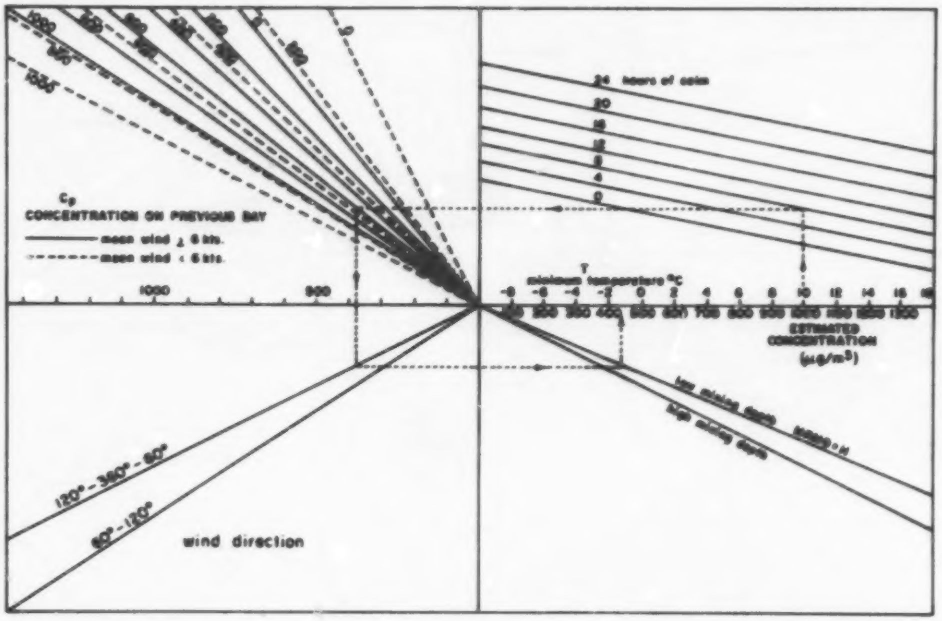

Tottenham
miles
1 2 3
1 2 3
km.
150
Dagenham
200
150
200
300
400
Westminster
400
300
River Thames
200
150
150
Woolwich
Wimbledon

Figure 12-8. Mean winter concentration for Inner London 1968-69-70.



24 hours of calm
20
16
12
8
4
0
C_p
CONCENTRATION ON PREVIOUS DAY
mean wind ≥ 6 kts.
mean wind < 6 kts.
1000
900
T
minimum temperature °C
-8 -6 -4 -2 0 2 4 6 8 10 12 14 16 18
100 200 300 400 500 600 700 800 900 1000 1100 1200 1300
ESTIMATED CONCENTRATION (μg/m³)
low mixing depth
high mixing depth
120° - 360° - 60°
60° - 120°
wind direction

Figure 12-9. Nomogram for SO_2 concentrations at the four sites in Inner London.

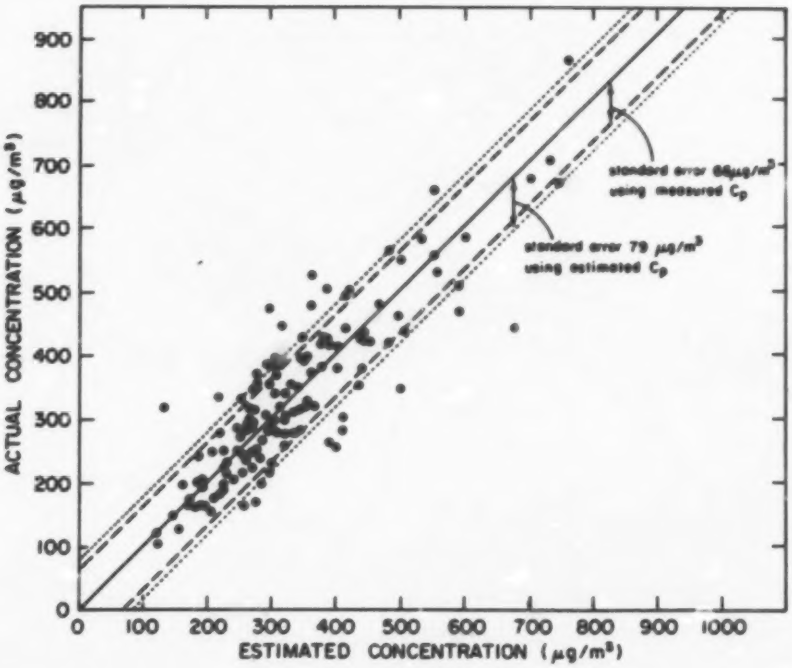


Figure 12-10. Results of the forecasting scheme of Part II.

Acknowledgements

The authors are grateful to the Warren Spring Laboratory for their helpful cooperation in the supply of their National Survey data, and to the following Councils who gave permission for the data from their areas to be used:

The Corporation of Manchester
The Royal Borough of Kensington and Chelsea
The Borough of Hackney
The London Borough of Lewisham
The Corporation of London
The Rural District Council of Epping and Ongar.

We also wish to acknowledge the very helpful work done by Mr. P. Bushby, a vacation student, who analysed the Manchester data. This paper is published by permission of the Director-General of the Meteorological Office.

References

Annual Reports of the National Survey, Warren Spring Laboratory.

The Royal College of Physicians, 1970: *Air Pollution and Health.* Pittman.

Weatherly, M. L. and Gooriah, B. D., 1970: National survey of smoke and sulphur dioxide: The greater London area. Warren spring Laboratory.

DISCUSSION

Gifford: You mentioned in your written presentation that you tried to avoid a numerical scheme because of the complexity that the computer required. Of course we have a scheme that is perfectly numerical which doesn't require a computer and since you had been good enough to include a whole rather long series of data we thought it would be fun to compare our scheme with yours. The results are summarized in the form of a comment to the preprint of your paper, which is presented after the discussion of the paper. Our scheme is very simple. In your notation it simply says that the concentration is proportional to the source strength divided by the wind speed. This might look like a box model to you, and in fact it is. This number here (C in Equation 1 of my comments on the preprinted paper), which can be expanded, depends rather weakly on the city size and on stability and if you are interested in seeing some other comparisons these were mentioned in the references. This model really works very well. For instance, it gives a correlation with the data that Joe Knox showed yesterday for carbon monoxide in San Francisco that is just about as good as the one from his model. Unfortunately, as you pointed out, you don't have the source strength to be entered into the model so we simply turn the formula around and use yesterday's source strength, which I call Q_0, to put into today's formula. So if 1 represents today and 0 represents tomorrow, our formula says that the concentration is given by the ratio of today's vs. tomorrow's wind speed, times the existing concentration level. Your model gives a 0.9 (rounded off) correlation; our model gives a 0.7 correlation. Your simplified model is 0.81 or perhaps it is 0.84, I don't remember. Anyway, persistence was low, 0.55. I put in confidence limits to indicate the fact that all of these models beat persistence. I discovered in going through the data that our model had a serious tendency to over predict. It being so simple, it's very subject, for instance, to error for low wind speed. When V_1 is 1 meter per second, things got pretty bad. So I tried an arbitrary correction, namely, I put a modified model which took the square root of the wind ratio and found this gave a correlation of 0.76. I would expect that these correlations, whichever one you choose to use, would repeat, since ours is essentially an *a priori* model. This I admit was somewhat suggested by our experience with your data, but I would expect that ours would repeat, that yours would come down a little bit, and that probably Manchester was more representative of the sort of effect that you are likely to get with it, and would guess that in all likelihood the best way to improve predictions of this item would be if you went to work on the source strength term. In fact I tried to incorporate some information and this is the question: Do you not think some information on the source spatial and temporal variability janitor functions and so on would improve the model?

I made the point in my prepared comment on your written paper that I thought that you had used both of your winters to develop this model and you've now told us that you only used one winter so I have to that extent more confidence. Just the same, yours is a very high correlation. If it holds up anywhere near that value it is going to be an awfully tough forecast to beat.

Smith: I am very pleased to see this and I would certainly agree that wind speed is important in this sense as you saw in the nomograms I showed you. Wind speed obviously was playing a very important role in that both in the number of hours of calm and in the influence of the previous day's concentration. I also agree that if one could somehow get a better understanding of the Q distribution, the source distribution in London, one might very easily improve the situation. The only hesitation I have in this is that both Dr. Gifford's and my correlations depend on using meteorological data which has really been measured, its real data. Whereas the forecaster's problem is he has to actually forecast this data which is not an easy task and I am quite sure the correlations will drop in practice quite significantly when the forecaster has to face the music and try to forecast the wind speeds and temperatures for the following period. And perhaps there comes the level in this when it's not worth going to great effort and expense to get a Q distribution when you know fairly well that even if you had the most perfect Q values the meteorology would let you down and you'll still not get a very accurate forecast.

Comment by F. A. Gifford on the Paper, "The Prediction of High Concentrations of Sulphur Dioxide in London and Manchester Air," by F.B. Smith and G.H. Jeffrey.

I would like first to record my agreement with Dr. Smith's remarks on the need for a simple approach, guided by the data, in air pollution forecasting. And also to second his remarks concerning the limited effect that mixing depth has on urban air concentrations, as a rule.

The authors approach this problem using the empirical techniques of objective weather forecasting and one can only applaud both their methodology and the workmanlike result that they obtain. They mention as alternatives to their empirical approach, numerical models and physical models, rejecting the former because they require a large computer facility. One seemingly valid numerical urban air pollution model does not however require such a computer facility, namely the simple ATDL model described by my colleague Dr. Hanna and me in a series of papers. (See all the references cited.)

It is of some interest to compare our model with the present results, particularly as Smith and Jeffrey have included a fairly long series of London SO_2 air concentration data, together with related meteorological data, the developmental data for their model. Application of our simple model to the present data is similar to the applications discussed in the first four references cited. Chemical SO_2 removal, which could be included using the scheme suggested by Hanna (1972), will not specifically be taken into account. Then the simple model gives, in the notation of Smith and Jeffrey,

$$C = cQ/\bar{v} \qquad (1)$$

where Q is source strength and c is a dimensionless parameter. Unfortunately no data on Q are included. To apply our model to forecasting London SO_2 concentration, we have to use today's value of C as a measure of Q, i.e., where subscript zero means today. Then

$$Q_0 = \bar{v}_0 C_0 / c \qquad (2)$$

is the prediction of our model.

$$C_1 = (\bar{v}_0 / \bar{v}_1) C_0 \qquad (3)$$

Table I displays the results of this comparison, in the form of correlation coefficients (with 95% confidence limits) between predicted concentration values from Equation 3, using the data in Smith and Jeffrey's Table II. In this table, "ATDL-modified" refers to a second prediction using the following modification of Equation 3:

$$C_1' = (\bar{v}_0 / \bar{v}_1)^{1/2} C_0 \qquad (4)$$

This entirely arbitrary modification attempts to account for the lack of data on source strength variability. Variation of Q with several of the meteorological parameters of Table II is probable. For instance, higher winds in winter are likely to be correlated with higher domestic fuel consumption. The modification to our simple model is an arbitrary attempt to take this into account.

Table I - Correlations of concentration predictions by various models with 140 measurements of London 24-hour SO_2 concentrations.

Model	Correlation Coefficient	95% Limits
Met. Office	.87	.91-.82
ATDL	.70	.79-.60
ATDL-Modified	.76	.82-.68
Met. Office-Simplified	.81	.86-.74
Persistence	.55	.66-.42

Several conclusions can be drawn from Table I. None of the other correlation coefficients fall within the confidence limits of persistence, so in this sense all methods "beat" persistence. Of these methods "Met. Office" correlates best with the data sample. But remember that this is an empirical method, developed from the given data sample. The "ATDL" correlation can on the other hand be considered a genuine test and should repeat on independent London data or on data from other cities. "ATDL-Modified" incorporates an *a priori* hypothesis and should also hold up. However in all honesty it has to be pointed out that the data suggested the hypothesis.

More important, it seems safe to suggest that there would be no inherent difficulty in proposing further empirical modifications to "ATDL" to bring it up to the level of "Met. Office," based on the data sample. "ATDL-Modified" is already quite competitive. But notice the following implication of that fact. "ATDL" says that only Q, v, and c can vary. The last varies with stability. To whatever extent Q, the source strength pattern, is a variable factor, "Met. Office" takes this into account indirectly, through correlations of Q with meteorological factors. If you believe "ATDL" on the other hand, it seems that the most sensible way to improve SO_2 forecasts would be to include the best available estimates of Q(x,y,t).

References

Gifford, F. A., and Hanna, S. R., 1971: Modeling urban air pollution. *Atmos. Environ. 6.*

Gifford, F. A., and Hanna, S. R., 1971: Urban air pollution modeling. Proceedings 2nd Int. Clean Air Congress, Washington, D. C., Academic Press, pp. 1146-1151.

Gifford, F. A., 1972: Application of a simple urban pollution model. Proceedings of Conference on Urban Environment and Second Conference on Biometeorology, Philadelphia, Pa., American Meteorological Society, pp. 62-63.

Hanna, S. R., 1971: Simple methods of calculating dispersion from urban area sources. *J. Air Pollution Control Association. 21:* 774-777.

Hanna, S. R., 1972: A simple model for the analysis of photochemical smog. Proceedings of Conference on Urban Environment and Second Conference on Biometeorology, Philadelphia, Pa., American Meteorological Society, pp. 120-123.

13. FITTING CURVES TO URBAN SUSPENDED PARTICULATE DATA

DAVID A. LYNN

Department of Statistics
Harvard University
Cambridge, Massachusetts

Introduction

The effort described here is a comparison of a number of theoretical frequency distributions with respect to their ability to fit bodies of urban air quality data. I have used almost exclusively some quite extensive suspended particulate data from Philadelphia, and so the work, and of course the conclusions, are to this extent limited and in a sense preliminary. It is presented at this meeting in particular, principally in the hope that it might interest others in actually trying out various distributions on their data.

The reaction of people in the air pollution field to the mention of this topic is often one of surprise. The question is usually believed to have been long settled by the use of the lognormal distribution and its parameters, the geometric mean and geometric standard deviation. It is certainly true that the lognormal is widely accepted in this role. I think we have all observed that air quality data is typically positively skewed, so that the normal distribution doesn't fit at all, and then taken the lognormal as the simplest positively-skewed alternative, without ever really considering alternative distributions very deeply. (Phinney and Newman (1972) is a recent example.) We need to remember that the decision to use the lognormal distribution was made mainly by the staff of the National Air Sampling Network back in the 1950's when computer-handling of data was almost unheard of. At that time, it was a major accomplishment to achieve enough computer processing to publish the NASN data summary reports, let alone to do anything very lengthy or complicated.

Now, however, the situation is much different. Not only do most sizable agencies and installations have computerized data processing, but we are beginning to be asked for increasingly sophisticated statistical inferences from our air quality data. This will likely become increasingly true as the typical pollution levels in our cities decline to the general vicinity of the National Ambient Air Quality Standards. It is my view that as we in the field of data analysis begin to provide more sophisticated methodologies to be used for these inferences, we should re-examine one of our most basic assumptions. In a sense, the increasing importance of gaseous pollutants during the 1960's has already forced this re-examination on us somewhat, and the results are not completely consistent—we have to use arithmetic means with sulfur dioxide data in order to avoid the embarrassment of too many logs of zero.

We might also consider briefly, by ways of further motivation, just what we have, and might in the future do, with a distributional form once we've chosen it, and how we verify our choice. When the NASN first introduced the lognormal, they used it in two ways. They included the geometric parameters in their data summaries, and they used logarithmic spacing of tallying intervals in the frequency distribution program that calculated the familiar summary line of percentiles. More recently, however, statistical applications have included significance tests, confidence intervals (Hunt (1972)), and extreme value statistics (Singpurwalla (1972)). In the future we may need inferences about spatial distributions, decision-theoretic-inference procedures, and so on, all of which involve making distributional assumptions of some sort. We might note for the sake of completeness that in some fields, notably actuarial science, some of the jobs done years ago with fitted theoretical curves have more recently been approached with smoothed empirical distributions. In fact at least one of these smoothing techniques has been applied to air quality data, specifically NASN particulate data (Spirtas and Levin (1970)). The reasons we might prefer a theoretical model to a set of smoothed data are the availability of probability theory, or at least more probability theory, the ability to use somewhat smaller data sets, and the ability to follow trends in the value of a parameter over time.

Distributions and Fitting Methods

Our purpose here is to compare the several theoretical distributions under consideration by actually fitting each to a number of sets of data, seeing which typically fits best, or which fits best in some overall sense, if indeed any do. The data used are suspended particulate data gathered at three sites in Philadelphia by the City control agency. The data have been gathered daily for many years. Here I have used data from 1960 to 1968, comprising 25 annual data sets in all—nine years from three stations with two sets missing.

There are a number of ways to determine the parameter values that will fit a specified distribution to a set of observed data. The use of the word "determine" rather than "estimate" the parameter values is deliberate. For most of the data under consideration we have essentially complete data for the entire year, so we consider the year's frequency distribution as "known", and view the problem as fitting a distribution with known parameters rather than as estimating the parameters from a sample. Of the various methods available, I have used the "method of moments," which selects the specific distribution out of the given family that has the same moments as our observed distribution, in each case equating as many moments as there are parameters in the algebraic formulation of the distribution. The first two moments are the mean and the variance, so this is of course just what we routinely do when we calculate the sample mean and standard deviation and then use a normal distribution with that mean and standard deviation, similar to what we would do with the two geometric parameters. Used in this way, the mean and variance of a normal distribution are effective as location and scale parameters. That is, changes in the mean are equivalent to changing the location of the distribution back and forth along the axis, while changes in scale, such as changes in the size of the units. Beyond these simple additive and multiplicative variations, however, a two-parameter distribution has no flexibility left to change its shape.

We can, however, have more than two parameters, and can determine their values by equating more than two moments from the observed distribution to the algebraic expressions for their theoretical counterparts. In common practice, four moments are the most used, because the sensitivity of the higher-order moments to small sampling errors is very great. The third and fourth moments, when used are commonly used in modified forms rather than in their raw numerical form. If we let μ_2, μ_3, and μ_4 be the variance and 3rd and 4th moments respectively, we commonly deal instead with the coefficients

$$\beta_1 = \mu_3^2 / \mu_2^3 \quad \text{and}$$

$$\beta_2 = \mu_4 / \mu_2^2 \tag{1}$$

The division by the proper power of the variance makes β_1 and β_2 small dimensionless numbers, while squaring μ_3 makes β_1 independent of its sign. These two constants (β_1 and β_2) are called the coefficients of skewness and kurtosis, respectively. Figures 1 and 2 give some indication of how they operate to measure the shape of a distribution. The point to note here is that every β_1, β_2 pair represents a different possible distribution shape, if the theoretical distributional form we are using has enough parameters to make use of them. This is of course not to say that any observed distribution with the same mean, standard deviation, skewness, and kurtosis will be the same. We have reduced some 300-plus data points down to four numbers, and have given up some information in the process, but we've given up less than if we dealt with only two parameters. The first few columns in Table I present the mean, standard deviation, and coefficients of skewness and kurtosis for the various data sets.

To pass from general considerations to the specific distributions used here, let's begin briefly with the lognormal. We have used not only the ordinary lognormal, here called the two-parameter or 2-p lognormal, but a three-parameter version which includes as the third parameter a location parameter, additive in the $\mu g/m^3$ scale before the logs are taken. The density, y, written in terms of the parameters GM3 and GSD3, and θ, is given in Equation 2, with GM3 and GSD3 denoting the parameters analogous to the geometric mean and standard deviation. The 2-p lognormal is obviously just a special case of the more general form. It doesn't really have a true location parameter, because it's tied to the fixed origin of the coordinate system by the nature of the log function. As the geometric mean varies along the axis, it does take the bulk of the density with it, but the shape also changes, as illustrated in Figure 3. The third parameter, here θ permits the distribution to slide along the axis independent of changes in its general size and shape.

$$y = \frac{1}{\sqrt{2\pi}\,(X-\theta)\ \ln \mathrm{GSD3}}\ e^{-1/2\left[\frac{\ln(X-\theta) - \ln \mathrm{GM3}}{\ln \mathrm{GSD3}}\right]^2} \qquad (2)$$

The other three-parameter distribution was the Gamma distribution with density

$$y = \frac{\left[\frac{X-\gamma}{\beta}\right]^{\alpha-1} e^{-\left[\frac{X-\gamma}{\beta}\right]}}{\beta\Gamma(\alpha)} \qquad \text{for}\ \begin{matrix}\alpha > 0\\ \beta > 0\\ X > \gamma\end{matrix} \qquad (3)$$

As one can see in the term $\frac{x-\sigma}{\beta}$, the parameters σ and β are location and scale parameters, respectively; α is a shape parameter that is determined from the skewness coefficient β_1. If the shape parameter is an integral multiple of 1/2, say n/2, and $\beta = 2$, $\sigma = 0$, the Gamma becomes a chi-square distribution with n degrees of freedom. As we see in Figure 4, the Gamma distribution can take on a variety of shapes with changes in α. For $\alpha < 1$, it becomes "J-shaped", that is, has an infinite ordinate at $x = \sigma$; as α gets very large, the shape approaches that of a normal distribution. Although it often isn't apparent in the graphs, the lower tail of the density curve does come down tangent to the axis.

The other distribution considered was the four-parameter Beta distribution or, as it turned out, the Beta distribution and another four-parameter distribution in different instances. The Beta is the closest there is to a common four-parameter distribution, although it's most commonly used in a two-parameter form. Its density,

$$y = \frac{\Gamma(p)\cdot\Gamma(q)}{\Gamma(p+q)} \times \frac{(X-A)^{m_1}\,(B-X)^{m_2}}{(B-A)^{p+q-1}} \qquad \begin{matrix}m_1 = p-1\\ m_2 = q-1\end{matrix} \qquad (4)$$

depends on two parameters, A and B, that are very clearly the minimum and maximum, and two others, p and q, that are symmetrical, representing a sort of relative contribution from the two ends. The shape of the distribution and its β_1 and β_2 coefficients are functions of the latter two parameters, and are quite flexible. The various distributions in Figures 1 and 2 are all Betas. The bounded nature of the Beta and the p-q symmetry make it a sort of continuous analogue of a binomial random variable and it is most often used as a Bayesian prior distribution for the binomial parameters p and q, with the max and min parameters set at 0 and 1.

Actually, the Beta distribution is specifically a Type I member of the system of theoretical density curves developed by Karl Pearson many years ago. Pearson's system is an attempt to provide a theoretical density function for every point in the β_1, β_2 plane, that is, for every possible combination of the skewness and kurtosis coefficients. As we see in Figure 5, his system consists of three "main types", I, IV, and VI, and a number of "transition types". The three main types are four-parameter distributions, representing the areas in Figure 5, while the transition curves are represented by lines and points at the boundaries of the areas. The three-parameter curves, such as the Gamma distribution, which is Pearson's Type III curve, appear as lines in the plane, while the two-parameter Normal distribution is only a point. Thus if one is fitting Pearson curves to data, the correct main type to be used depends on the values of β_1 and β_2—if we try to fit the wrong one, we get imaginary numbers during the calculation process. As it turned out, the Philadelphia data used here, as often as not, fell outside the range of the Beta distribution, we also included Type VI when needed. There developed no need for Type IV, though there very well might have. The Type VI density function is most simply written in terms of an arbitrary origin, Equation 5. A more complicated formulation, with the data expressed in their normal way, is more logical (Eq. 6). In either case, it's apparent why Type VI is not often used.

We note in closing that the lognormal densities, though not a part of Pearson's system, can still be represented by a line in the β_1-β_2 plane,

TYPE VI DENSITIES

$$y = \frac{(A^*)^{q_1-q_2-1}\,\Gamma(q_1)}{\Gamma(q_1-q_2-1)\,\Gamma(q_2+1)} \times \frac{(X^*-A^*)^{q_2}}{(X^*)^{q_1}} \qquad (5)$$

with X^* and A^* referred to an origin at

$$\mu - \frac{(q_1-1)\,A^*}{q_1-q_2-2}$$

with origin at 0 μg/m^3,

$$y = \frac{(q_2+1)^{q_2}(q_1-q_2-2)^{q_1-q_2}\,\Gamma(q_1)}{X_0(q_1-1)^{q_1}\,\Gamma(q_1-q_2-1)\,\Gamma(q_2+1)} \times \frac{\left[1+\frac{X-\mu}{A_2}\right]^{q_2}}{\left[1+\frac{X-\mu}{A_1}\right]^{q_1}} \qquad (6)$$

$$\text{where } A_1 = \frac{X_0(q_1-1)}{(q_1-1)-(q_2+1)} \quad \text{and } A_2 = \frac{X_0(q_2+1)}{(q_1-1)-(q_2+1)}$$

as seen in Figure 5.

Johnson and Kotz (1970) and Elderton and Johnson (1969) are most useful references for various distributions.

Results

There are several possible methods of quantifying how well the theoretical curves fit the observed distributions. None of them are obviously superior. As is apparent from any of the histograms, the discrepancies between the curves and the observed distributions will include a good bit of non-smoothness in the observed data as well as any lack of fit. It's probably equally apparent that we don't really want to fit every bump in the observed data, but rather want to eliminate the bumps. One can smooth the data in some mathematical way, or one can merely aggregate it with the distributions into intervals of various widths, with fewer and broader classes presenting the smoother-looking appearance. As it turned out with the present data, it didn't seem to affect the relative judgment among the distributions, so the simpler approach was chosen. The observed data were first tallied into classes 5μg/m^3 in width, and the frequencies were compared with estimates calculated from the theoretical curves by a simple quadrature technique. The frequencies and the estimates were then aggregated into classes of 10 μg/m^3 and then 20 μg/m^3 in width.

In each case, the comparison criterion used was the sum of the absolute differences between the observed and the expected or estimated frequencies. The magnitude of this criterion varied strikingly with the level of aggregation, and relatively little from one distribution to another, as shown in Table II. Only very rarely did the different levels of aggregation affect the relative fits of the different distributions, and then only trivially. For comparison purposes, a normal distribution was fitted in each case, and the ratio of the criterion above to the similar value for the normal was considered as a possible criterion, in the hope it might be largely independent of the level of aggregation. It didn't prove to be, however; the symmetric normal fits the skewed data so badly that the effect of the poor fit dominated the effect of the roughness of the observed data. This comparison did, however, provide an interesting way of viewing the process of choosing a distribution. With the value of the criterion for the normal being 100%, the choice of any of the skewed distributions would reduce this to 50-70%, while the distributions typically differed among themselves by less than 10%, and rarely by more than 30%.

In fact, when we look at the results in Table III, it's apparent that there is frequently very little choice among the distributions. The figures in the table are the values of the absolute difference criterion at the 20 $\mu g/m^3$ level of aggregation; to provide a simple summary, averages for each station and for the whole group are provided. It's clear, first of all, that the normal distribution doesn't fit, which is of course no surprise. It is also apparent that the two lognormal distributions have noticeably lower values than the Gamma and the four-parameter Pearson. It is not really clear that the differences between the two distributions in each pair are real because position changes of 1 to 2 units or so did sometimes occur just in changing from one to another of the various fit criteria.

Because the data used here are quite a narrow sampling from the many monitoring sites (with their possibly-different distributional shapes), we probably want to attach relatively little weight to choosing an overall "winner" among the distributions, and rather consider just why the various distributions react the way they do to the different data sets. In Table I are tabulated the fitted values of the parameters of the various distributions, and with this information in conjunction with Table III and some figures we'll consider the performance of the various distributions.

First we consider the Pearson curves, the Gamma distribution and the two types of four-parameter curves. They are usually closer together than they are close to the performance of the two lognormals. But except for a very few instances, they aren't too far from the lognormals, at least in comparison with the normal. The extremely bad instances are the cases where one or both of them becomes J-shaped. Recalling that in the β_1-β_2 plane, the Gamma represents the dividing line between the Type I and Type VI areas, we can view the Gamma as an approximation to either. The equation of the Type III, or Gamma line, is $2\beta_2 = 6+3\beta_1$, so we would expect the approximation to be better, the closer β_1 and β_2 for our distribution are to that line.

In fact, the Gamma and the Type I or VI are quite close for most of the data sets (or station-years), and those that are farthest apart are in fact those farthest from the line. Considering the data sets where the Gamma and the four-parameter are the most different, we find that those where the four-parameter is better (1-60, 1-65, 2-65, 3-65) are all in the Type VI area, while those where the Gamma distribution does better (1-68, 2-62, 2-67, 2-68, 3-68) are all in the Type I area. In fact, with a single exception, closer examination shows this to be true for all the data sets. Except for 3-66, the

Gamma always does better than the Type I and worse than the Type VI, though in some cases the differences are trivial. It's not immediately clear why this should be, and given the fact that the years 1965 and 1968 occur so often in the lists above, it is likely nothing more than the fact that a year's meteorology may give a distinct shape to the distributions at the several stations.

Figure 6 presents histograms and the fitted curves for two examples of cases where the Pearson Type VI does much better than the Gamma. Figure 7 presents two cases where the Gamma does better than the Pearson Type I. In Figure 6 it's clear why the Gamma does poorly - it makes a much larger peak than necessary, and then underestimates on the downslope of the upper side of the distibution. In contrast, in Figure 7, where sharper peaks are needed, the Pearson Type I tends to exaggerate the peak or, as in the 2-62 case, even to go off to an infinite ordinate. In these cases, the curves are almost identical on the upper side of the peak. On the lower side, the Gamma does better with its moderate peak, though neither does really well. Interestingly, the factor which controls these cases is hardly evident in the resulting curves. Those where the Type VI does better than the Gamma are typically those with high maxima, or long tails, and those where the Gamma does better than the Pearson Type I are those with quite short tails--those with few or no values over 400 $\mu g/m^3$. The 1-65 case in Figure 6 is a good illustration. The Gamma comes more steeply down the upper side of the curve in order to throw more probability mass out into the tail, and in order to do so, builds itself a higher peak, losing good fit not only in the peak but on the downslope as well.

Before considering the two lognormal distributions more thoroughly, it might be of value to consider when and why the two pairs of distributions differ more or less. Although the two Pearson distributions are poorer overall here, they have a number of possibly-useful properties, and ought not be discarded outright. Clearly, the situation where they differ most is in those high-skewness cases where the Pearson curves peak wildly or become J-shaped. No matter what the skewness, the lognormals retain their zero ordinate at the lower boundary, and hence fit much better.

The two pairs, however, do sometimes differ fairly markedly even when the Pearson curves don't have infinite ordinates. Data set 3-68 in Figure 8 is an example where the two lognormals are better, and set 3-66 is Figure 9 one where the two Pearson curves are better. These two sets of data illustrate fairly well the typical result in those situations when there is a noticeable difference among the distributions, typically the cases where the distributions' tails are moderate-well out into the 300-400 $\mu g/m^3$ range, but not into the 600 $\mu g/m^3$ range. On the upper side of the distribution, there is very little difference among the several distributions. To accomplish this, the two Pearson curves typically have somewhat sharper peaks than the two lognormals, and more abrupt slopes on the lower tails. Thus if the observed distribution has a clean, sharp peak as does 3-66, the Pearson curves fit better, while if it has a broader peak as in 3-68 or 2-64, the lognormals fit better. Actually, in these data, the former is a rarity; the

3-66 data set is quite unusual, being the only case where the Pearson Type I is the best of the four, and the only exception to the pattern of the Gamma doing better than the Type I, regardless of the rank.

If we consider the differences between Figures 8 and 9, we would rather expect that those observed distributions with peaks, in some sense, midway between would produce roughly the same fit with all the curves. This is precisely what happens. Figure 10 shows two data sets for which the results are quite close, with only the best and worst of the four distributions plotted. Their peaks do seem midway, being a little more blocky on the high than on the low side. In addition, there is less noise than in some of the distributions, leaving less opportunity for one of the distributions to appear better or worse, fortuitously. In fact, the 3-62 data set can be considered the best in this sense, because by a clear margin, it had the lowest overall value of the fit criterion.

There are several such data sets where all four distributions fit about equally well; in about half of all the sets, three of the four could be described as essentially tied. The balance of cases on either side of this center is far from even, though; there are several data sets with square peaks, while only one (3-66) with a sharp peak. It is largely this predominance of relatively square, blocky peaks (low kurtosis) that gives the lognormals their overall advantage, at least among these data. And at this point it might be prudent to point out that sharper peaks often seem to go with overall lower concentration levels. This is the case here, and since this is relatively high data (especially for the early years), we might expect somewhat different overall results with other data.

We now turn to a closer look at the differences and similarities between the two lognormals. The first general observations are most readily seen in the list of parameters in Table I, comparing the geometric mean and standard deviation with their analogues in the three-parameter model, labelled GM3 and GSD3. The location of the distributions remains strikingly constant, the sum of the threshold location parameter θ and GSD3 is rarely more than 3 or 4 $\mu g/m^3$ from the two-parameter geometric mean. The parameter θ takes on both positive and negative values, many near zero but with a few sizeable cases in each direction. We also note that when the value of θ is negative, GSD3 is less than GSD, and vice versa, in rough proportion to the magnitude of θ.

In terms of performance, we've already noted that the two-parameter lognormal does overall slightly better than the three-parameter, and in fact does the best of all four distributions tried. This is, of course, in conflict with what might be expected, since with only two parameters, it should have the least flexibility. We've seen why this happens, though. The three-parameter lognormal, the Gamma, and the four-parameter Pearson are in order of increasing flexibility, but they use this flexibility primarily to adjust to the upper tail of the distribution. While the 3-p lognormal can't go off to a wild peak or an infinite ordinate even remotely as easily as the Pearson curves, it can do so more readily than the 2-p lognormal, as in Figure 11. This permits the 2-p distribution to "win" by default, even though it fails badly to fit the blocky peak. Thus the success of the lognormals is a bit left-handed, as it were. They succeed not because they are sensitive to the bulk of the data but simply because they are insentitive to the upper tail, and are tied down at the lower tail.

Figure 12 also illustrates this peaking tendency, but here the result is the opposite; the observed distribution (2-62) has a very sharp peak, and the flexibility of the 3-p lognormal reaches it a little better, though neither does really well. It is also possible to associate this tendency toward high peaks and long tails with the value of θ. The larger positive values are associated with those that have shorter peaks and shorter tails than their 2-p counterparts, such as 3-66 in Figure 9.

An attempt to summarize is not as difficult as it first appears. The overall level of performance is the reverse of what would be expected on the basis of the number of parameters involved in the distributions, largely because of the sensitivity of the distributions to the extreme upper tail. As a consequence, square, blocky distributions can't be fit well unless they have a very short tail. And I should add as a closing footnote, though this isn't the place to discuss it, that this sensitivity to the extreme upper tail is characteristic of fitting by the method of moments. Thus the first place to look for improvements is likely in less touchy, though possibly less simple, fitting methods.

TABLE I(a)
Fitted Parameter Values

		MOMENTS				LOGNORMALS				
Sta/Yr		Mean	Std. Dev.	β_1	β_2	Geo. Mean	GSD	GM3	GSD3	θ
1	1960	169	71	1.75	5.85	156	1.50	158	1.49	-2
	1961	167	74	2.84	7.28	153	1.51	128	1.63	23
	1962	156	58	1.35	4.85	147	1.43	146	1.43	0
	1963	167	80	5.81	12.69	153	1.51	95	1.87	57
	1964	174	69	1.65	5.47	162	1.46	157	1.48	4
	1965	166	68	2.88	8.77	154	1.46	117	1.63	34
	1966	174	69	1.54	5.45	162	1.46	163	1.46	-1
	1967	152	69	1.99	6.08	139	1.53	142	1.53	-3
	1968	136	55	1.43	4.83	125	1.49	136	1.44	-10
2	1960	142	58	2.19	6.61	133	1.46	113	1.55	18
	1961	137	54	2.08	6.07	128	1.45	109	1.54	17
	1962	136	55	2.96	6.88	128	1.43	93	1.64	31
	1963									
	1964	137	56	1.71	5.36	127	1.47	125	1.49	2
	1965	141	56	2.56	7.90	132	1.44	102	1.60	27
	1966	147	61	2.52	7.18	136	1.46	111	1.59	23
	1967	127	57	1.85	5.38	116	1.53	123	1.51	-6
	1968	118	49	1.61	5.05	109	1.49	114	1.47	-5
3	1960	151	67	2.18	6.60	138	1.52	133	1.55	4
	1961	153	70	2.14	5.93	139	1.53	140	1.54	-1
	1962	138	53	1.33	4.99	129	1.45	135	1.43	-6
	1963	152	65	1.70	5.37	140	1.50	147	1.48	-7
	1964									
	1965	134	65	3.87	11.76	122	1.55	95	1.72	24
	1966	138	67	1.61	4.92	124	1.59	155	1.47	-30
	1967	131	64	1.71	5.44	117	1.60	143	1.49	-24
	1968	117	56	1.78	5.37	105	1.60	123	1.50	-17
9	1968	123	59	1.42	4.57	110	1.62	147	1.44	-34
11	1968	152	74	1.58	4.87	137	1.59	171	1.47	-32

TABLE I(b)

Fitted Parameter Values

		GAMMA				PEARSON TYPES I & VI				
Sta/Yr		α	β	γ	Criterion	X. (VI)	A	m_1/q_1	m_2/q_2	B(I)
1	1960	2.29	47.1	61	4.04	1553	50	44.03	2.06	1731
	1961	1.41	62.7	79	87.89	30908	78	500.5	0.43	
	1962	2.96	33.6	57	-3.82		64	1.32	38.83	1731
	1963	0.69	96.7	101	5.46	1396	93	19.93	-0.04	
	1964	2.42	44.4	67	-104.		67	1.39	1001.	44973
	1965	1.39	58.0	85	1.30	227	43	16.57	4.44	
	1966	2.59	42.7	63	5.74	2346	56	66.60	2.14	
	1967	2.01	48.6	55	11.92	4493	52	101.84	1.21	
	1968	2.79	33.1	43	-2.31		55	0.87	19.86	1034
2	1960	1.82	42.7	65	4.06	1134	56	36.07	1.47	
	1961	1.92	38.9	62	-24.70		63	0.84	184.1	7569
	1962	1.35	47.5	72	-3.49		80	-0.06	13.87	1040
	1963									
	1964	2.33	36.4	52	-4.48		58	0.86	35.97	1703
	1965	1.56	45.0	71	1.51	272	44	17.79	3.42	
	1966	1.58	48.1	70	3.89	1064	61	30.87	1.21	
	1967	2.15	39.0	43	-2.49		54	0.43	15.99	996
	1968	2.48	31.4	40	-2.32		50	0.64	17.37	876
3	1960	1.83	49.7	60	3.85	1249	49	34.73	1.53	
	1961	1.87	51.3	57	-4.31		65	0.44	26.74	1839
	1962	3.00	30.5	46	-48.48		47	1.94	575.1	17954
	1963	2.36	42.4	52	-5.24		58	0.94	43.49	2302
	1964									
	1965	1.03	63.4	69	1.01	26	-11	47.32	38.32	
	1966	2.48	42.6	32	-1.71		50	0.43	11.48	903
	1967	2.34	41.7	33	-6.99		38	1.01	59.26	2901
	1968	2.24	37.6	33	-3.19		41	0.63	22.92	1228
9	1968	2.82	35.4	23	-1.31		44	6.86	0.44	922
11	1968	2.53	46.2	35	-1.65		56	9.73	0.44	1122

TABLE II
Typical Summary Tables With Criterion At 5, 10, and 20$\mu g/m^3$ Class Widths

STATION 2, 1964

	K	Normal	2-P Log	3-P Log	Pearson	Gamma
	SUM	360.5632	363.0000	363.0000	363.0208	362.8891
5$\mu g/m^3$	DIFF	146.4098	105.9059	106.0737	120.8522	114.4267
	REL	1.0000	0.7234	0.7245	0.8254	0.7816
10$\mu g/m^3$	DIFF	125.9653	66.1073	67.0827	86.2695	80.6849
	REL	1.0000	0.5248	0.5325	0.6849	0.6405
20$\mu g/m^3$	DIFF	125.6084	38.8479	39.1794	62.3205	53.4492
	REL	1.0000	0.3093	0.3119	0.4961	0.4255

STATION 3, 1961

	K	Normal	2-P Log	3-P Log	Pearson	Gamma
	SUM	320.3271	324.9995	324.9990	325.7625	324.5805
5$\mu g/m^3$	DIFF	168.2665	109.4671	110.7158	121.5336	114.4060
	REL	1.0000	0.6506	0.6580	0.7223	0.6799
10$\mu g/m^3$	DIFF	148.1114	79.7891	81.1142	101.8797	90.2807
	REL	1.0000	0.5387	0.5477	0.6879	0.6095
20$\mu g/m^3$	DIFF	126.3808	39.1026	39.2929	56.9197	53.0919
	REL	1.0000	0.3094	0.3109	0.4504	0.4201

TABLE III

Summary of Total Absolute Deviations (20μg/m^3 classes)

		Normal Dist.	Lognormals		Pearson	
			2-P	3-P	4-P	Gamma
Station	1960	109.8	43.4	43.2*	45.6	54.0
1	1961	135.1	56.6*	62.6	109.3	110.1
	1962	129.9	45.0*	46.7	49.7	49.4
	1963	159.4	60.8*	93.1	131.6	210.6
	1964	119.5	59.3*	60.3	70.9	70.5
	1965	129.1	59.0	55.9	55.1*	84.8
	1966	129.0	52.2	52.4	48.3*	48.4
	1967	131.0	51.2*	52.5	60.6	63.9
	1968	120.3	60.2	59.4*	86.6	69.3
	Average	129.2	54.2	58.5	73.1	84.6
Station	1960	114.2	46.6	39.5	36.4*	39.1
2	1961	125.7	41.6*	46.1	69.3	68.1
	1962	177.2	82.2	64.4*	105.7	68.2
	1963					
	1964	125.6	38.8*	39.2	62.3	53.4
	1965	119.5	43.5*	46.5	48.3	63.5
	1966	143.5	43.3	35.8*	52.1	67.0
	1967	143.7	59.8*	63.6	75.6	62.9
	1968	134.5	60.4*	61.3	80.9	68.9
	Average	135.5	52.0	49.6	66.3	61.4
Station	1960	108.6	47.3*	47.8	57.2	66.7
3	1961	126.4	39.1*	39.3	56.9	53.1
	1962	106.0	27.7*	28.1	29.6	29.1
	1963	122.9	56.4	57.8	48.0	45.1*
	1964					
	1965	123.8	56.0	54.6*	57.4	89.4
	1966	134.9	58.6	66.8	50.0*	51.7
	1967	131.2	53.7	53.8	52.2	49.7*
	1968	136.9	61.1*	65.3	82.5	73.1
	Average	123.8	50.0	51.7	54.2	57.2
Station	1968	94.0	60.3	58.9*	72.8	60.7
9						
Station	1968	60.2	31.6*	36.0	36.9	32.0
11	Average	125.6	51.7	53.0	64.1	66.8

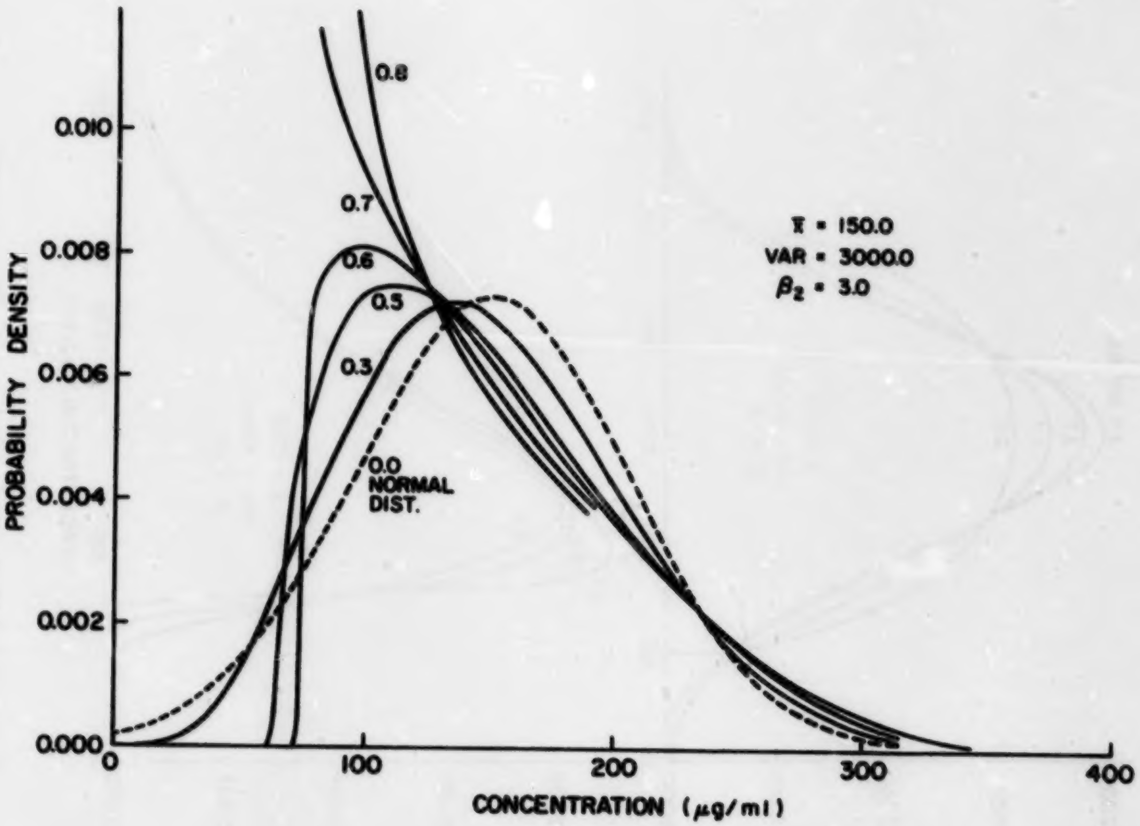


Figure 13-1. Effect of skewness β_1 on distribution shape.

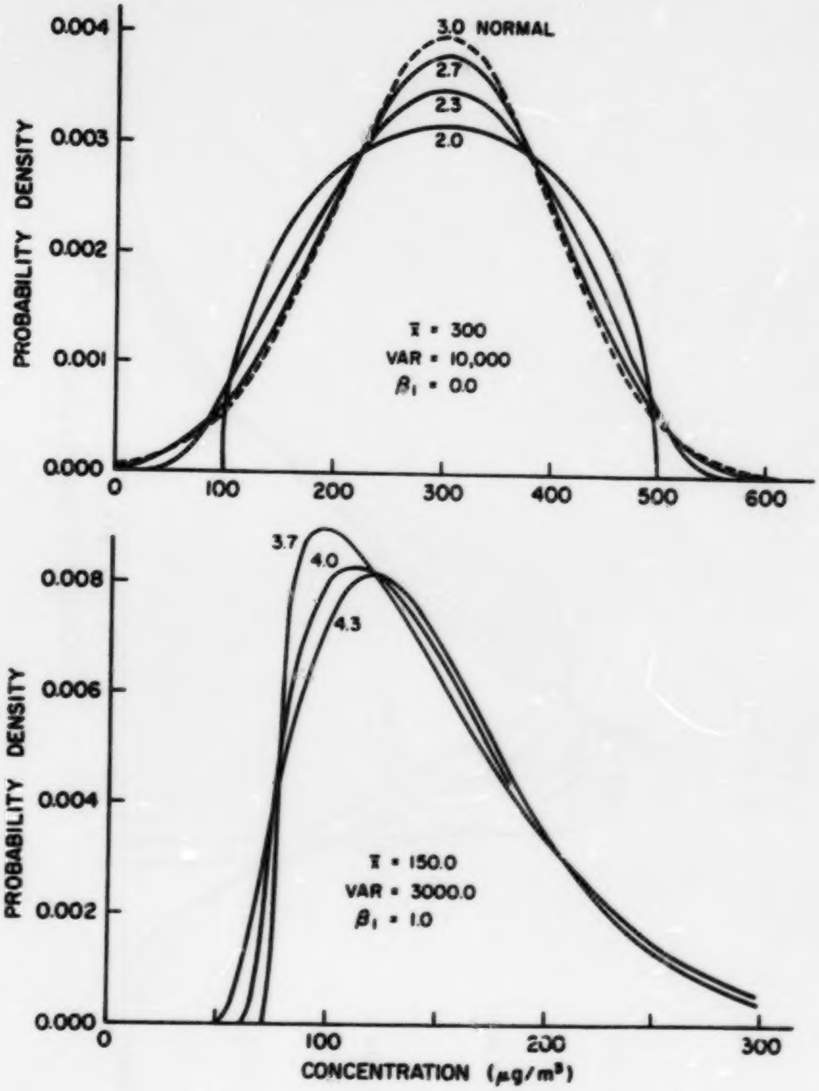


Figure 13.2. Effect of kurtosis β_2 on distribution shape.

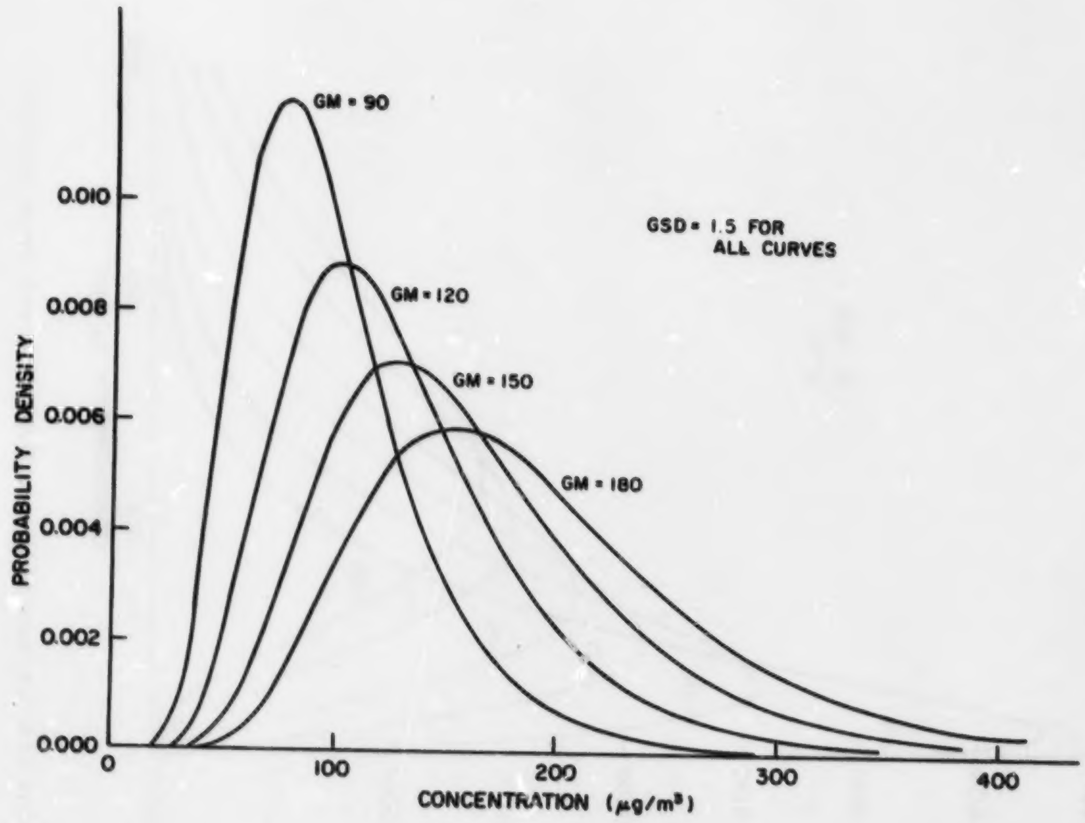


Figure 13.3. Two-parameter lognormal density curves.

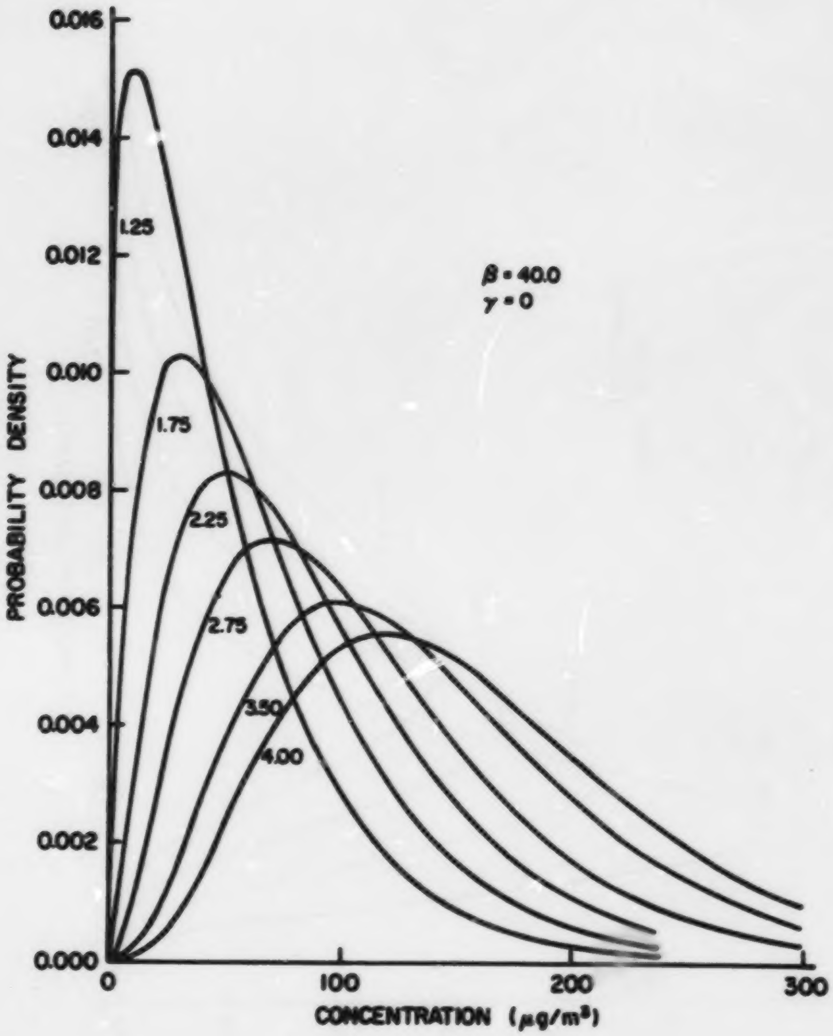


Figure 13-4. Variation of Gamma distribution with shape parameter *α*.

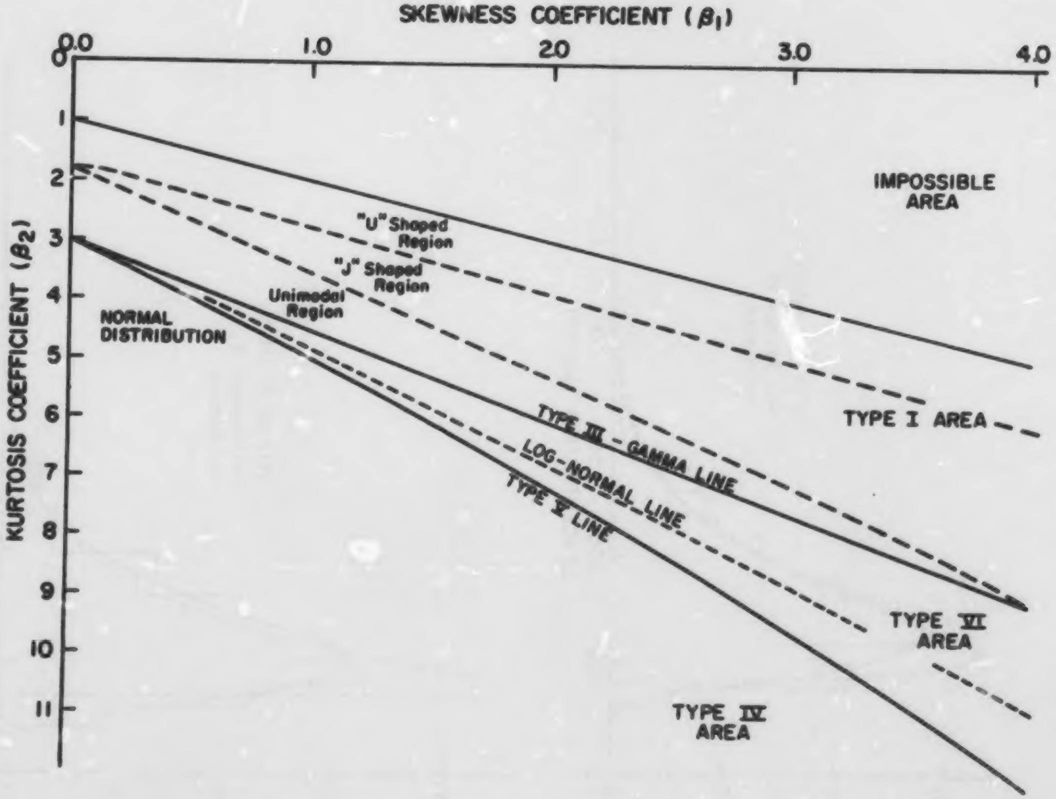


Figure 13-5. β_1, β_2 Plane with Pearson system.

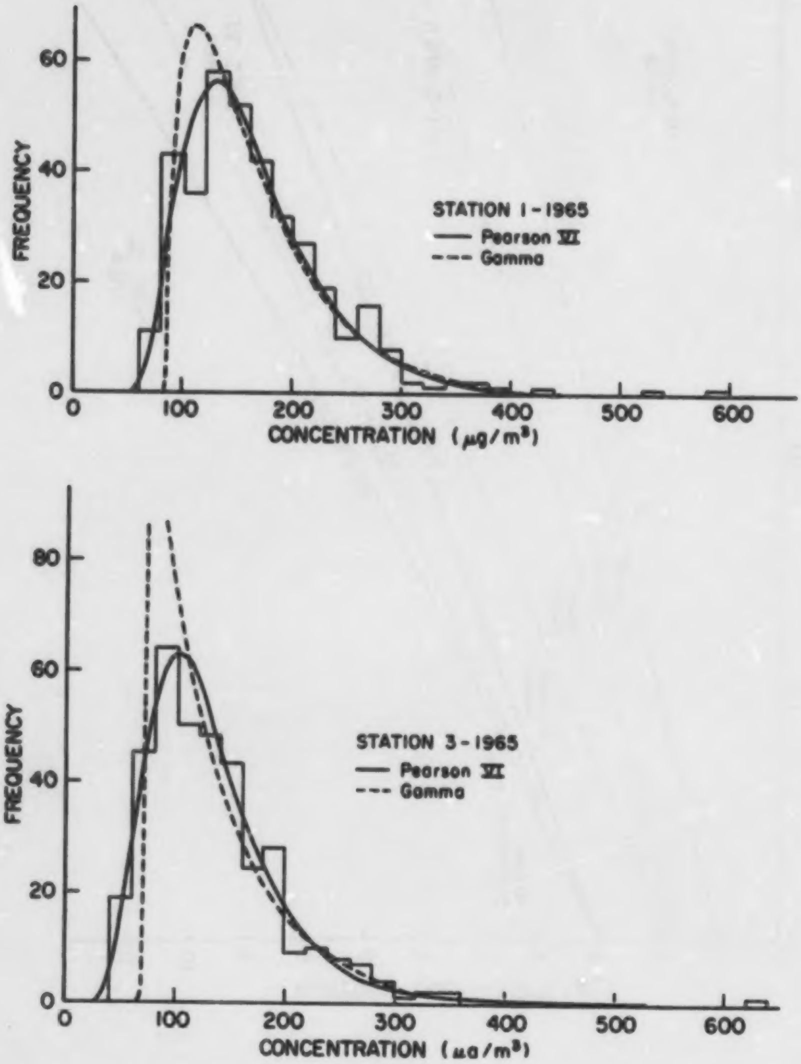


Figure 13-6. Pearson VI does better than gamma.

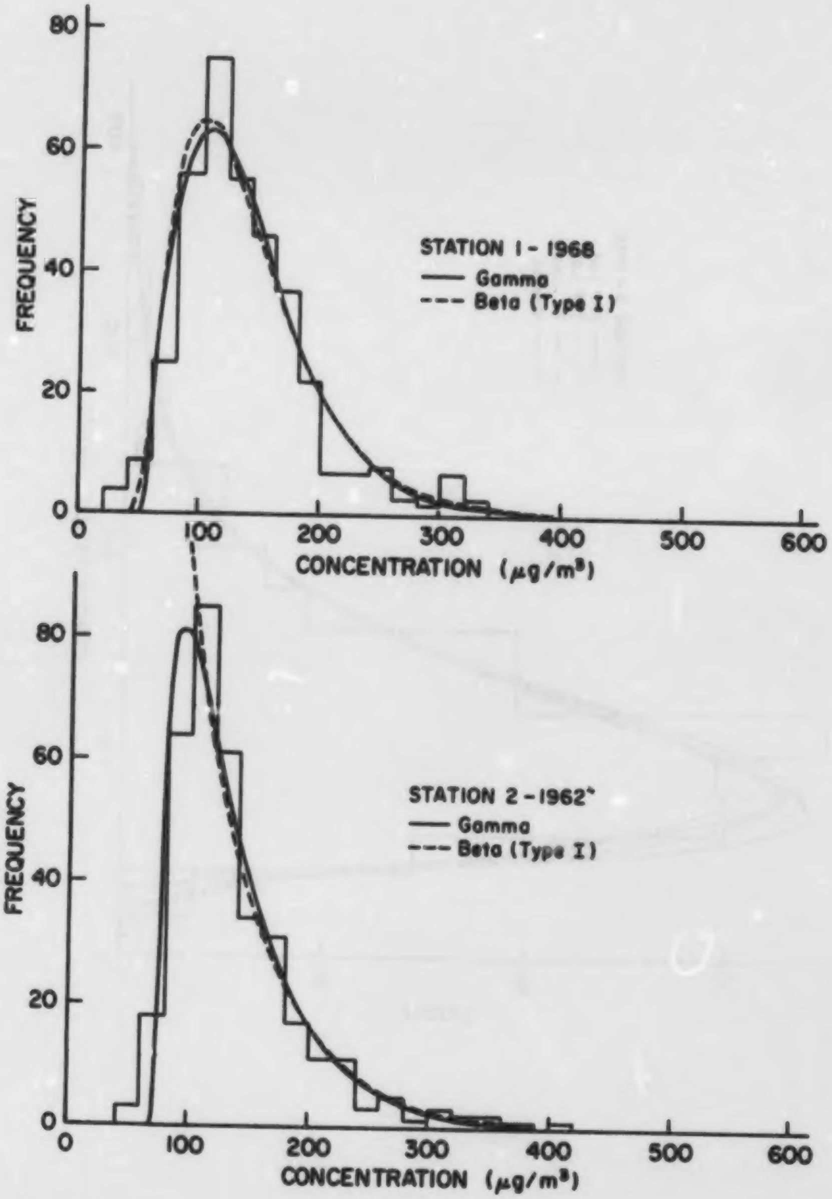


Figure 13-7. Gamma does better than Pearson I.

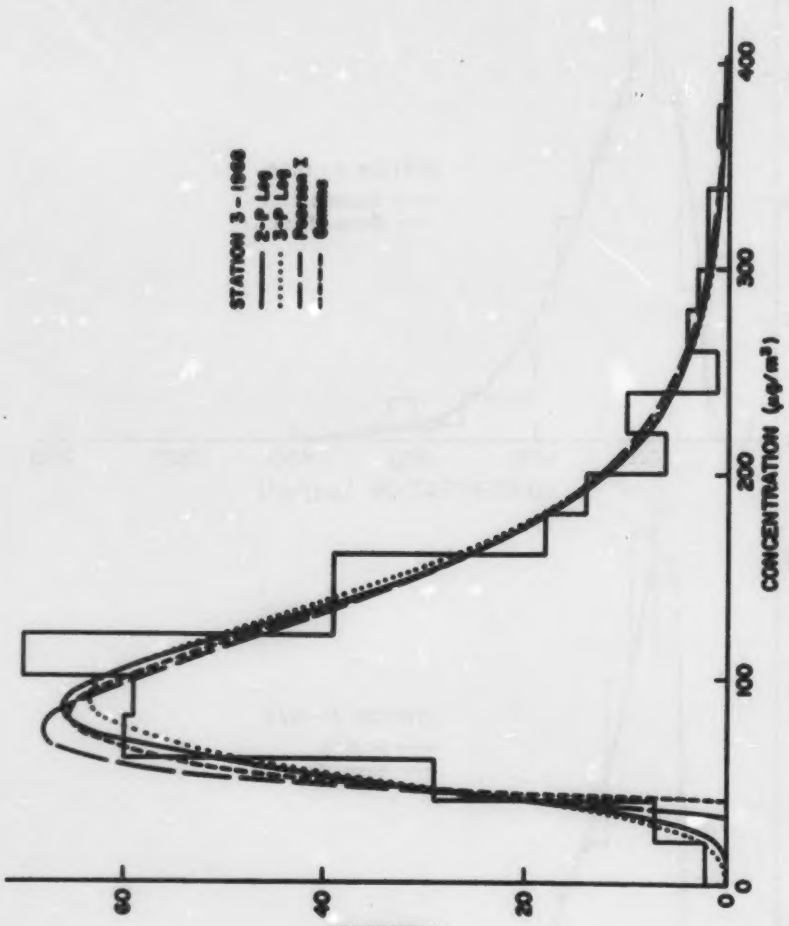


Figure 13-8. Logs better than Pearson curves.

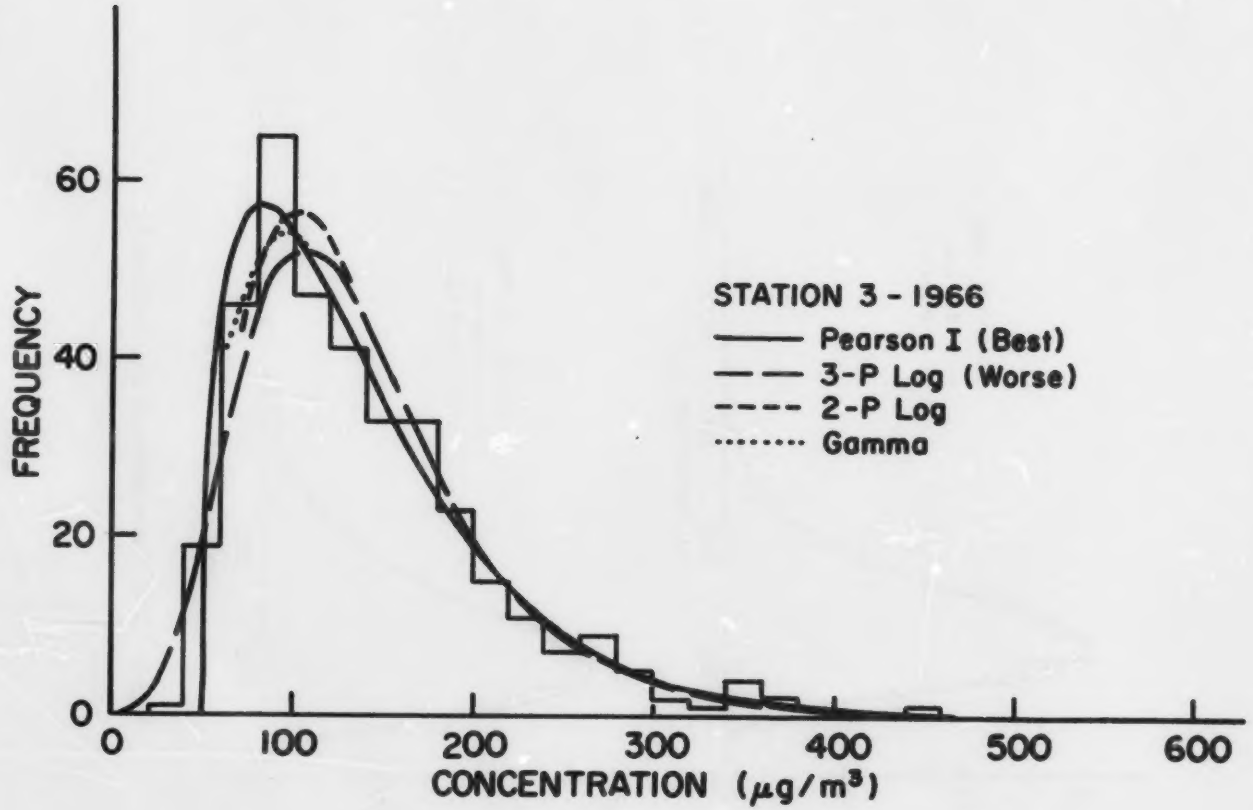


Figure 13-9. Cases where Pearson curves do better.

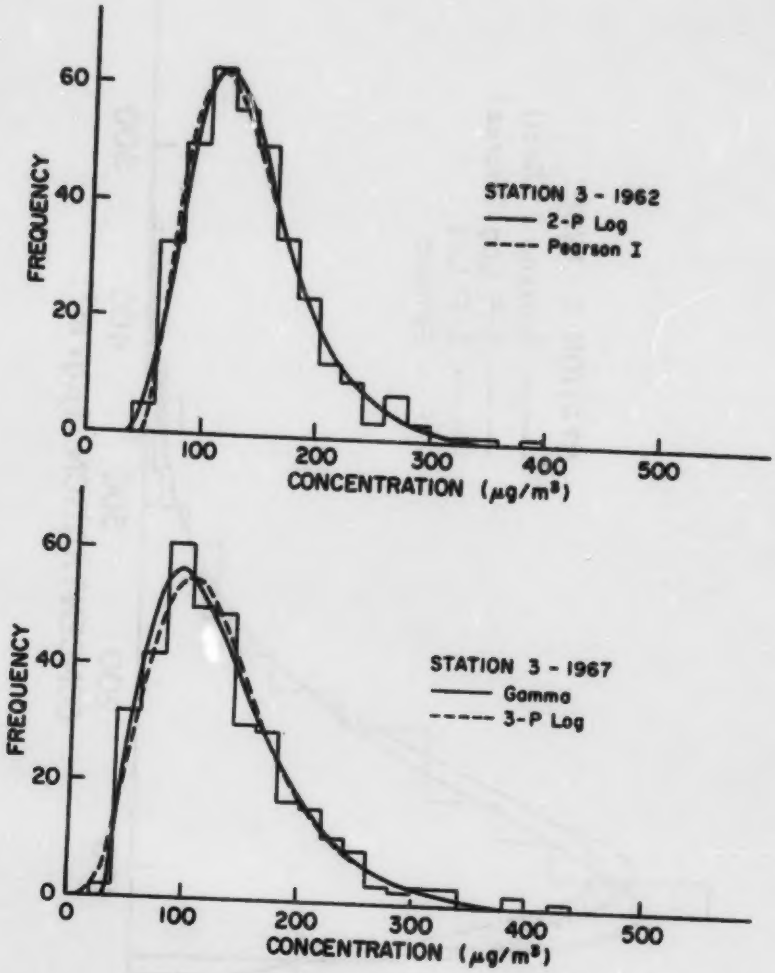


Figure 13-10. Cases where all distributions fit.

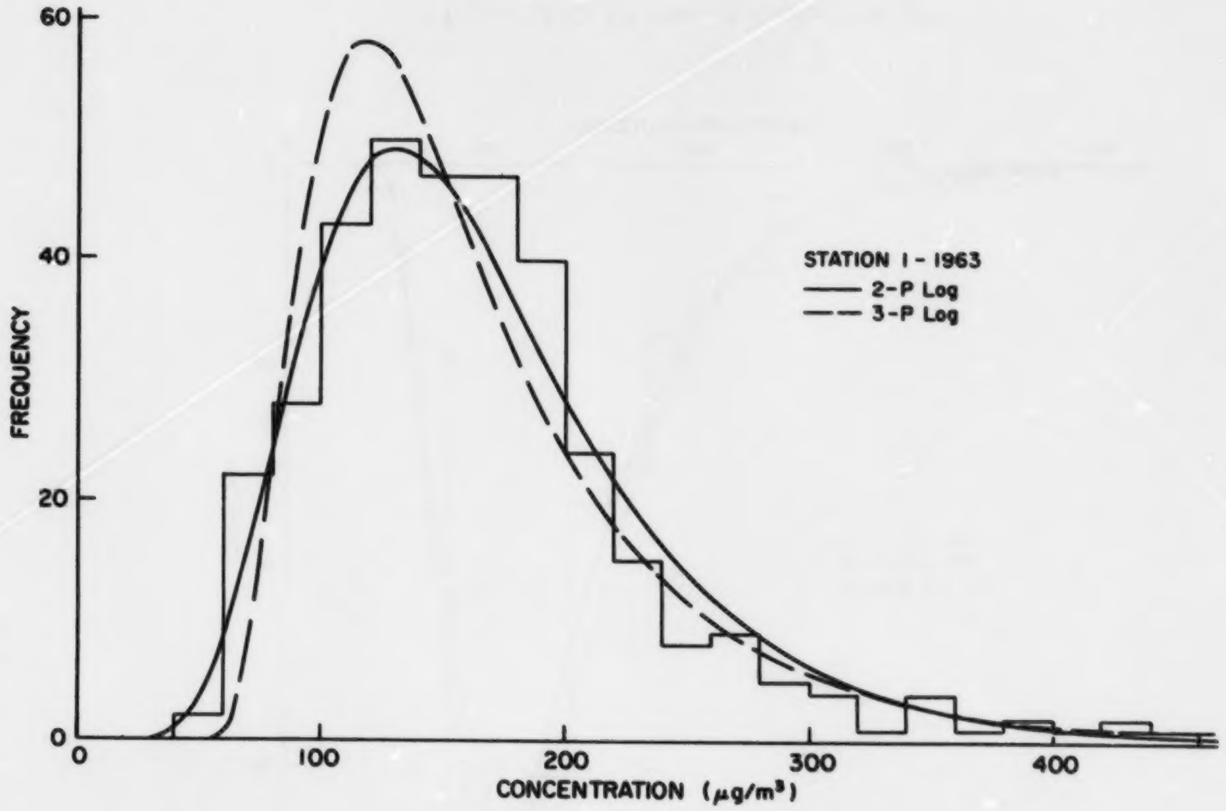


Figure 13-11. 2-P lognormal better than 3-P.

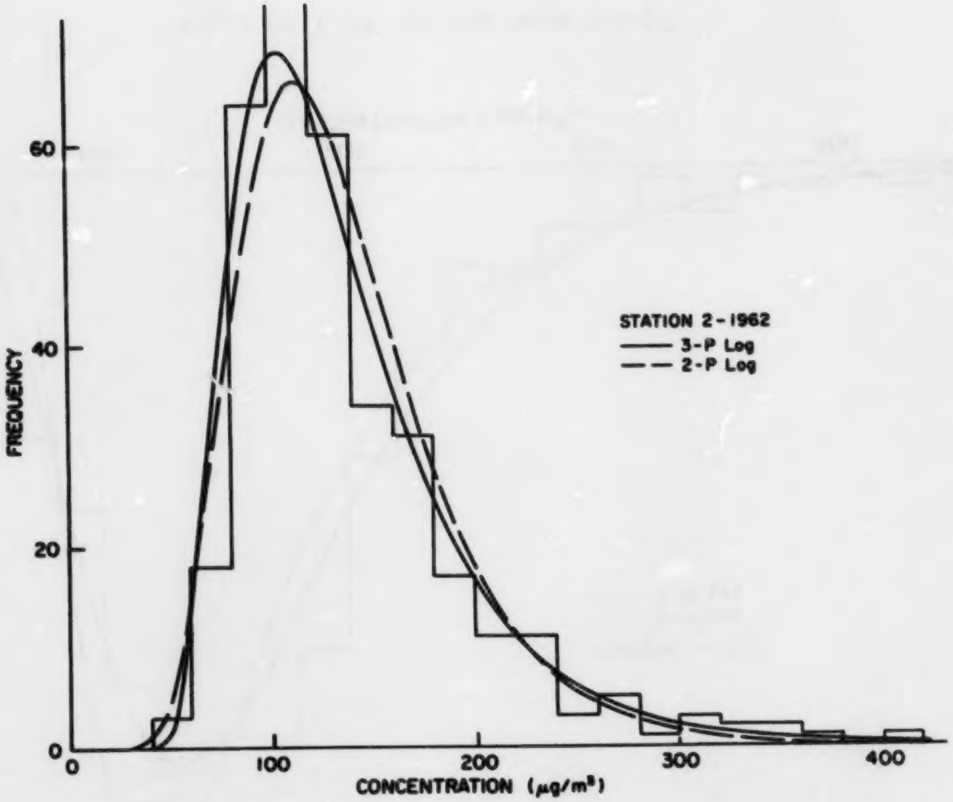


Figure 13-12. 3-P lognormal better than 2-P.

Acknowledgement

My appreciation is gratefully extended to the Harvard Department of Statistics for their support of the computing necessary for this effort, and to the Philadelphia Department of Public Health for permission to use their data.

References

Elderton, W. P., and Johnson, N. L., 1969: *Systems of Frequency Curves,* Cambridge University Press.

Hunt, W. F., Jr., 1972: The Precision Associated with the Sampling Frequency of Log-Normally Distributed Air Pollution Measurements. *J. Air Pollution Control Association. 22*: 687.

Johnson, N. L., and Kotz, S., 1970: *Continuous Univariate Distributions, Vol. I and II.* Houghton-Mifflin.

Phinney, D. E., and Newman, J. E., 1972: The precision associated with the sampling frequencies of total particulate at Indianapolis, Indiana. *J. Air Pollution Control Association. 22:* 692.

Singpurwalla, N. D., 1972: Extreme Values from a Lognormal Law with Applications to Air Pollution Problems, *Technometrics. 14:* 703.

Spirtas, R. and Levin, H. J., 1970: Characteristics of Particulate Patterns 1957-1966. PHS, NAPCA, Publication AP-61.

DISCUSSION

Larsen: As you mentioned in the beginning, Dave, there are several distributions that could be used for expressing skewed data. Now that we are considering controlling air pollution, there is one simple characteristic where the lognormal is handy for the man in the regional office. For instance, if we have a lognormal distribution of SO_2 concentration and we use a control plan that reduces all sources by 90%, then we expect the new distribution to be parallel and just down one log cycle on the plot. So the lognormal provides a very simple way to predict what is going to happen to a future distribution based on the present distribution. Also it gives a physical feeling of what might happen if, for instance only tall sources were reduced, maybe all the power plants in the region, and all the low, area sources were not controlled. You might tend to chop off the highest observed concentrations to give a new distribution with a shallow slope. Or if only all the low sources were controlled, a steeper slope would be expected.

Lynn: There's not much I would say; I certainly agree that nothing much fancier than the lognormal is ever going to be what we call handy for that type of field calculation. I think that the place of more complicated distributions is in fitting large bodies of data as entire bodies of data, rather than making assumptions about control strategies. And of course the case when one wants to do some mathematics which the lognormal may not be amenable to, or just that nobody has done that particular derivation yet. I hesitated there in saying whether another distribution would ever be of use in this field because I think the extreme value distribution can be handy and do well if one is interested only in predictions about the maximum levels.

Hershfield: We have been talking a great deal about the 2-parameter lognormal distribution but we haven't mentioned some important characteristics of the distribution. We have a distribution which is skewed to the right. It is outlier prone. If you plot the coefficient of skew (C_s) on the vertical axis versus the coefficient of variation (C_v) on the horizontal axis, the lognormal distribution has the relationship given by the equation, $C_s = 3C_v + C_v^3$. Some mention has been made of the extreme-value or Fisher-Tippett type I (Gumbel) distribution. Gumbel, in his work postulates a coefficient of skew 1.139 and at C_v equal to 37%; estimates from both the lognormal and the Gumbel distributions are equal.

14. A PROPOSED AMBIENT AIR QUALITY SAMPLING STRATEGY AND METHODOLOGY FOR THE DESIGN OF SURVEILLANCE NETWORKS

JOSEPH R. VISALLI AND DAVID L. BRENCHLEY

Environmental Engineering Laboratory
School of Civil Engineering
Purdue University
West Lafayette, Indiana

and
HOWARD REIQUAM

Battelle Memorial Institute
Columbus, Ohio

Introduction

Determination of the concentration of a specific pollutant in the ambient air over a defined area is a complex problem because of spatial and temporal variations in air quality. Many factors such as local topographical differences, atmospheric chemistry phenomena, fluctuating emission rates, non-stationary sources, and varying meteorological conditions contribute to these variations. This space-time dependence complicates immensely the acquisition of statistically accurate estimates of true air quality. This in turn confounds the interpretation and utility of the data.

The Federal Clean Air Act of 1970 requires the monitoring of ambient air for various pollutants. Well defined and vigorous federal directives have been issued indicating which pollutants are to be monitored, and the preferred reference testing methods and procedures for measuring these pollutants (anon. (1971)). In contrast, the guidelines for the establishment of air quality surveillance networks are essentially of a subjective nature as implied by the following quotation: "Experience and technical judgment are essential for determining the number and location of sampling sites because adequate mathematical models or other models have not been formulated" (E.P.A. (1971)). The importance and high costs associated with ambient air monitoring demand that mathematically rigorous methods for determining the numbers and placement of sampling instrumentation be developed, such that the precision associated with the test methods and procedures is of equivalent magnitude to that of the actual data collection. It is of little value to take precise measurements in a subjective manner. The result is high cost data, having poor reliability and statistical validity, and of little utilitarian value in determining compliance with Federal standards.

The Problem

Federal air quality regulations essentially limit the concentration of specific pollutants to certain levels at all locations within defined areas (Anon. (1971)).

Basically then, the problem is reduced to one of finding the point or points within the defined areas at which the concentration of the specific pollutant is the greatest. Noting however, that an infinite number of points exist in the defined space, and that the concentration levels vary both spatially and temporally, the deterministic approach of finding the points where the concentration is the highest is a practical impossibility. Instead, a probabilistic approach is indicated. A random sample of pollutant concentration within the space over the defined area should be taken. This sample must take into account both space and time factors. From the resultant distribution, an estimate of the probability that a certain level of pollutant concentration has been exceeded can be made. Two problems arise as a result of choosing this approach. One is concerned with the mechanics of designing and conducting a simple random sample with respect to an infinite population that is a function of time and space. The other pertains to a strategy and methodology that will enable an unbiased estimate of the mean and variance of pollutant concentration to be calculated, with pre-determined accuracy and confidence in the statistical sense. An approach to the first problem can be devised by taking into account the nature of the turbulence regime at the microscale level. To accomplish the latter requires an advance estimate of the sample size needed for assurance that the desired degree of precision is attained (Cochran (1963)). This in turn requires an advance estimate of the variance of the random variable under consideration. With this advance information, a random sample of sufficient size can be conducted, such that an accurate and precise estimate of the air pollutant concentration distribution over the defined area can be made. The probabilistic question of whether air quality standards have been exceeded for the sampled area can then be asked. More important is that with such a procedure, the question can be answered with a mathematically determined degree of confidence.

The Strategy and Methodology

Strictly speaking, federal standards can be interpreted to include not just ground level concentrations, but concentrations as a function of the vertical coordinates up to the depth of the mixing layer. To accomplish such a task would require, as will soon be evident, a virtual armada of airplanes or helicopters, or a remote sensing method. The primary purpose of establishing federal standards is to protect public health and welfare. Thus we feel it can be reasonably argued that the interests of public health and welfare will be adequately served if pollutant concentration is monitored with respect to both space and time at people (ground) level only. Dispensing of the vertical aspect brings the problem into the realm of practicality as far as economics, and the satisfaction of sampling criteria (to be discussed later) are concerned. The scope of the problem, even with this significant deletion, remains however, beyond practical consideration at this point. The results of the Nashville study (Stalker et al. (1962)) revealed that at least 245 sampling stations, approximately four per square mile, would be required to estimate the daily mean concentrations of sulfur dioxide for the whole city with 95% confidence of ±20% accuracy. Establishment of this many monitoring stations is probably beyond the financial capabilities of most communities. The only alternative then, if statistical accuracy and precision are to be maintained, is to consider that only certain sub-areas within the whole bounded area require air monitoring. These sub-areas would be specific sectors of the whole bounded area, that by some subjective or objective criteria are judged to contain the highest levels of pollutant concentration.[1] Effectively then, we would be utilizing the statistical technique of stratification, but departing from the usual method of analysis. Ordinarily, sampling would be conducted in randomly selected stratified sub-areas, and an estimate of the whole area mean value for some variable calculated from these samples. In this approach, we purposely select the stratified sub-areas judged to contain the highest values of the variable under consideration, and attempt to establish from the sampling of these sub-areas an upper bound to the variable. We can then conclude that the remaining sub-areas, based on the developed criteria, will be enjoying better air quality (the variable) than that indicated by the established upper bound.[2] With this approach, the scope of the problem is within the limits of practical economic consideration, the legal aspect is satisfied, and the mathematical accuracy and precision that is desired can be attained.

---

[1] A certain amount of subjectivity is inherent in most engineering approaches or solutions to problems. This is due to the two classical compromises between mathematical requirements and technical capabilities, and between this resolution and available financial resources. It will become evident later, that the approach taken in the proposed methodology would theoretically eliminate all subjectivity if the imposed economic demands could be met. In any case, deciding what constitutes a critical area is a complex question that can be approached in any one of, or a combination of the following ways. It could be based on existing air quality isopleth data, simulation dispersion model predictions, attitudinal-behavioral studies of the citizenry, demographical-medical-specific area data, wind rose data in conjunction with industrial location distribution, or prior complaints. The point is that some criteria can be developed.

Consider one of the sub-areas discussed earlier. It is a defined area of constant (with time) topographical features. Assume a hypothetical situation where the emissions polluting this sub-area are of constant strength and flow rate. If the meteorological factors that influence the transport and dispersion of these pollutants are also assumed to remain constant with time, then the mean and variance of pollutant concentration across the sub-area will remain constant with time. Larsen (1971) indicates such a situation for a single point. One can extrapolate the concept to cover an area simply by taking the average of many single points over the area.

Consider now the hypothetical case where emission factors polluting this sub-area remain invariant with time, but the meteorological conditions are changing with time. During time periods when the meteorological conditions are similar, our previous argument of a constant mean and variance of pollutant concentration is valid. That is to say, varying meteorological conditions can be classified according to some finite scheme that attempts to categorize these varying conditions into specific meteorological regimes (e.g. Pasquill's scheme). If these regimes are described by the factors that influence transport and dispersion phenomena, then during periods when a specific regime is dominating the weather, a specific mean and variance of pollutant concentration will accompany it.

In reality, however, emission characteristics will vary with time in a manner which is very difficult to predict. Thus even during periods of "similar meteorology", the mean pollution concentrations will differ. However, since the variance is not directly proportional to the mean it is reasonable to hypothesize that the variance of pollutant concentration will remain constant with time during periods of "similar meteorology". Effectively we are hypothesizing that since the variance of a group of numbers depends only on their relative magnitudes, and not their absolute values, the variance of pollutant concentration is not dependent upon the mean pollution level. Hence the variance is independent of emission factors, and the original argument for the variance of pollutant concentration holds.[3] Thus, differing concentrations of pollutants will be dispersed over a given area in a similar manner during periods of "similar meteorology".

---

[2] Considering the extreme complexity of the problem, it is entirely possible that points in the sub-areas not sampled could violate the standards. Nothing can be done about this problem using this methodology. Thus it is important to consider this aspect when determining the sampling criteria.

[3] The strategy and methodology developed in this paper does not consider the effect of non-stationary emission sources such as automobiles. Hence this approach is applicable only to those pollutants (such as SO_2) emitted primarily from stationary sources.

This hypothesis parallels closely the statistical concepts of homogeneity and stationarity as applied to horizontal turbulence levels over an area. Stationarity essentially means that the statistical characteristics which describe the frequency distribution of the horizontal turbulent eddies remain constant with time. Thus the mean, variance, and the rest of the statistical moments that characterize the frequency distribution remain constant with changes in time. Homogeneity implies that the frequency distributions of these turbulent eddies are the same throughout the area under consideration. Dispersion models derived from both gradient transport and statistical theory must employ these concepts in order to solve their respective equations. It is well known that the statistical properties of turbulence vary in the vertical direction. However, horizontal turbulence characteristics at constant height do fulfill to a certain extent, the statistical concepts of homogeneity and stationarity provided that major changes in the meteorological regime do not occur (U.S. Atomic Energy Commission (1968)).[4] The first hypothesis contained in our methodology states that a stationary process will occur, provided that no changes in the meteorological characteristics that govern the transport and dispersion phenomena occur. This is a less restrictive stationary process than the one postulated for the turbulence regime. We have proposed the concept that the variance of pollutant concentration, not the mean, will remain constant with time under specific conditions. In many stationary processes, assuming the mean of a random variable to remain constant with time is a questionable assumption. The stationary process we propose does not require the mean value to remain constant over time, and thus is less restrictive than the typical case.

Consider an infinitesimal section of a defined area. The variance of pollutant concentration will be very small, and in the limit will equal zero as the area approaches zero. As the section is increased in size, the variance will exhibit a tendency to increase according to some function toward the variance of the entire defined area. A decrease appears unlikely because the variance will decrease in magnitude only if observations tend to cluster about the actual mean. This indicates that the concentration gradient is tending toward zero. Experience demonstrates the reverse. There is a tendency for pollutants to accumulate in certain areas, creating large gradients when the whole area is considered. Thus it is hypothesized that an increasing area-variance relationship will exist. This is also in accordance with turbulent theory, which indicates that the statistical properties of turbulence in the horizontal will tend toward greater dissimilarity as larger areas are considered. Thus the variance between turbulent eddies will in all likelihood increase as larger areas are considered (U. S. Atomic

---

[4] It must be noted that these concepts are generally assumed to hold for flat plains and rolling countryside. Little is known about the effects of irregular terrain.

Energy Commission (1968)). As shown in Figure 1, knowledge of the variance-area relationship could be combined with cost data and available financial resources to maximize the area covered by the sampling strategy. Knowing what financial resources are available for sampling purposes and the cost per sample, we can easily determine the sample size we must work with. If we select a confidence interval (precision) and a margin of error (accuracy) that are deemed adequate for the purpose, then from the equation:

$$n = \left(\frac{t}{d}\right)^2 \text{ (variance)} = k_i \text{ (variance)}$$

where:

n = sample size, if the population size is very large
t = the abscissa of the normal curve that cuts off an area α at the tails
d = chosen margin of error
$1 - \frac{\alpha}{2}$ = confidence probability
$k_i = (t/d)^2$, where the value of t depends upon the sample size

we can determine an estimate of the variance that would be required to accomplish our sampling task. Knowing this and the variance-area size relationship, we can determine the size of the area we should cover in our random sample in order to achieve the prescribed precision and accuracy with the financial resources available. We obviously would like to maximize the area covered since this reduces the amount of subjectivity that has entered our strategy and methodology.

The Experiment

The correct method of actually obtaining the samples to be used in estimating the mean and variance of pollutant concentration is not at all obvious. To apply probability theory to the data with any degree of confidence, it is imperative that random sampling techniques be employed (Cochran (1963)). Keeping this in mind, one might propose that to overcome the spatial-temporal dependence noted earlier, continuous air monitors should be randomly scattered throughout the defined sub-area. This approach would be correct if the continuous monitors were not fixed at every randomly chosen location. Two reasons dictate that continuous monitors should be continually moved in random fashion, if accurate estimates of pollutant concentration and variance are to be obtained. First, the population (molecules of SO_2) is continually changing its spatial distribution with time, and its size (number of molecules) is also continually changing with time (molecules are added, deposited, transformed to other forms, remain with the area, etc.). Essentially then, a new random sample is needed for each new population. Secondly, and perhaps more important, the response of any monitoring instrument placed in an irregular topographical setting will be biased. This is due to specific effects of building geometry on diffusion parameters. If a monitor is fixed at one point—even if randomly located—the response of the instrument will include this bias on a continual basis with time. Effectively this states that no single point can be representative of a large area (Corn (1970)). This is especially true over relatively short periods of time such as a day. The Nashville study (Stalker et al. (1962)) demonstrated that relatively few instruments were needed to accurately estimate seasonal and yearly concentration levels over an area. This is a predictable result, as long term sampling tends to average out fluctuations and converge on the true mean. The study also revealed how inaccurate their results were when a few instruments were used to estimate 24 hour averages. This was due, in part, to the bias encountered from fixed point sampling. Thus, to continually "remove" or reduce the inaccuracy due to this time factor bias, one must continually relocate the monitors in a random manner with time. Randomized mobile monitoring is thus essential to proper sampling procedure in this case.

In accordance with this type of reasoning, we should also note that a restriction to the interpretation of data collected from continuous fixed point monitoring is indicated. If the purpose of the monitoring is merely to indicate trends at that specific location with time, continuous fixed station monitoring is certainly reasonable. Caution must be exercised however, if such data is to be used to determine *why* the trends occur over time. Changes in the topography, both local and in the vicinity of the continuous fixed point monitor (urban renewal, road construction, etc.), the addition of new sources, etc. can confound meaningful interpretation. It would be impossible to discern whether changes in air quality are due to an effective (or ineffective) air conservation management program, or merely due to changes in topography that alter the bias suggested previously. Problems of this nature would not arise if a mobilized monitoring approach was used. This is because the mobilized monitoring approach eliminates (or at least reduces) this bias.

Data analysis and prediction models

The utility of this whole approach is dependent upon an ability to predict accurate advance estimates of the variance. The choice of a suitable prediction model depends to a great extent upon the nature of variance data presented as a time series. Figure 2 depicts some of the possible results of graphically representing such data. All of these possibilities have distinct features that would influence the choice of a suitable stochastic prediction model. Figure 2A, for example, illustrates a process that incorporates features of both time stationarity and non-stationarity. This time series depicts discrete periods of stationarity that obviously change mean values in an almost instantaneous jump fashion. Figure 2B exhibits a similar tendency, but has a smoother transition between time stationarity intervals. Figure 2C shows a random fluctuation process. Figure 2D depicts a time series that is stationary in the wide sense.[5] Certainly other possibilities exist. Subsequent data analysis would reveal if the random variables are dependent or independent, and if long period, seasonal, or cyclic trends exist along with inherent random fluctuations.[6] A procedure outlining a general method of determining the nature of a particular set of data can be found in many texts on time series analysis (e.g. Box and Jenkins (1971)).

At this point we can only speculate (since data is not available) as to the nature of the time series. Since many possibilities exist, it is not feasible here to outline a data analysis for each situation. Instead, a single case where a time series that is in accordance with the hypotheses stated earlier will be considered. One possible approach to the data analysis will be outlined. The stress will be on simplicity.

Table I represents the type of data required for the approach outlined in this paper. The variance estimates and concomitant meteorological parameters are arranged in the sequence in which the data were taken. Each variance estimate is calculated from the average of simultaneous measurements taken over equal averaging times if discrete sampling techniques are employed, or the average of the measurements over equal time intervals if continuous monitors are utilized.[7] As shown in Figure 3, the variance can be plotted as a continuous function of time. Since the sequence does not appear to fluctuate in a totally random manner, and many atmospheric phenomena are known to be dependent processes (U. S. Atomic Energy Commission (1968)) it will be assumed that the variance measurements are dependent random variables. The data analysis should concentrate on achieving five objectives:

---

[5] The concept of stationarity in the wide sense essentially provides for a process where only up to a certain statistical moment is constant with time. A good possibility exists that many atmospheric processes exhibiting stationary features are actually stationary in the wide sense.

[6] The first hypothesis described in this paper provides for transient variations due to short term meteorological conditions. Cyclic trends due to diurnal effects and seasonal weather trends may also affect the time series. Long term trends, due perhaps to changing topology (urban renewal, etc.) or new emission sources, are also a possibility that should be considered.

[7] Averaging time is defined as some time interval over which the continuously varying trace of the variable may be represented by a constant (average) value. We note also that this averaging time (or the time interval if continuous monitors are used) should be as small as is technically possible to avoid the time bias discussed previously.

Table I. Data Sequence Required

V_1	V_i	V_n
X_{11}	X_{i1}	X_{n1}
	X_{ij}	
X_{1m}	X_{im}	X_{nm}

where:

V = spatial variance of pollutant concentration at time i

X_{ij} = recorded values of meteorological variables assumed to have a significant relationship to the variance

i = time period number

j = representative of a specific meteorological parameter

1. Determine periods of time stationarity over which the variance is homogeneous or constant.

2. Determine which meteorological parameters influence the changes in variance levels.

3. Characterize the meteorology during each period of stationarity according to some finite classification scheme based on the parameters found to influence variance changes. Test the hypothesis of "similar meteorology" based on this classification scheme.

4. Choose an algorithm that will predict advance estimates of the variance across the sub-area, given that the variance of the time period immediately preceding it is known, and the estimated meteorological parameters of the time period to the predicted are known.

5. Develop estimates of the length of the stationary time period for the predicted variance.

Keeping these objectives in mind, the data analysis *can* proceed as follows:

(a) Find the longest possible continuous time intervals over which the variance is constant in the statistical sense, i.e., no significant difference exists, or limit the standard deviation of the data to a certain percentage, etc. This will have the effect of breaking down the data into discrete step functions of time, where the variance will be constant over the interval. These resultant time intervals may not, as a consequence of the analysis, be of equal length, but should be as long as statistically possible. The overall result could look something like Figure 4.

(b) Run an analysis of variance or a regression analysis on all of the data to determine the meteorological variables that significantly influence the variance change.

(c) Based on the results of the analysis of variance (or the regression analysis), devise a classification scheme of a finite number of divisions or classes. The scheme should probably be both qualitative and quantitative in nature—perhaps similar to that proposed by Pasquill (1962). After each time interval is classified, we can proceed to test the hypothesis of "similar meteorology" by simple inference tests on the variance values of each class.

(d) Assuming that the hypothesis is not rejected, we are now in a position to propose a simple algorithm. This will predict an estimate of the variance in advance, given that we know what the variance was at the previous time interval, and have available an estimate of the values of the pertinent meteorological variables for the time interval to be predicted. This *can* be accomplished in the following manner:

(1) Breaking down the data into a finite number of classifications, say six (A,B,C,D,E,F), will result in six frequency distributions being developed (see Figure 5). The variance values for each class stem directly from re-arranging and grouping the data after classification. We thus have a mean value and a variance associated with each variance class. These distributions need not be similar. The only requirement for further analysis is that the variance of the variance distributions be finite.

(2) Choose a simple linear least-squares algorithm such as:

$$\hat{\nu}_{T_{i+1}} = E\left[V_{T_{i+1}}\right] + \frac{\mathrm{cov}\left[V_{T_i}\ V_{T_{i+1}}\right]}{\mathrm{var}\left[V_{T_i}\right]}\left[\nu_{T_i} - E\left(V_{T_i}\right)\right]$$

where:

$(\hat{\nu}_{T_{i+1}})$ = variance estimate for time period T_{i+1}

$E[V_{T_{i+1}}]$ = expected value of the distribution of the variance for time period T_{i+1}

$\mathrm{cov}[V_{T_i} V_{T_{i+1}}]$ = covariance of the random variables $V_{T_i} V_{T_{i+1}}$

$\mathrm{var}[V_{T_i}]$ = variance of the distribution of the variance for time period T_i

(ν_{T_i}) = variance for time period i

$E[V_{T_i}]$ = expected value of the distribution of the variance for time period T_i

We possess all the inputs to such an equation from the six frequency distributions. The

$$\mathrm{cov}\left[V_{T_i} V_{T_{i+1}}\right] = E\left[V_{T_i} V_{T_{i+1}}\right] - E\left[V_{T_i}\right] E\left[V_{T_{i+1}}\right]$$

The estimate for E $[V_{T_i} V_{T_{i+1}}]$ would be obtained from the bivariate distributions resulting from a Markov chain analysis of the sequence of order that the classification followed.

Time interval	1	2	3	4	5	etc.
Class	A	C	D	A	B	etc.

We can, as a result of this sequence determine the bivariate frequency distributions of AC, CD, DA, AB, etc. for all possible combinations of classifications to determine the cross-correlation function above.[8]

This whole analysis thus far depends upon the availability of advance weather forecasts. These forecasts must include estimates of the parameters found by the analysis of variance to be pertinent to variance changes. It must be noted that in the event such forecasted parameters are not available, we can determine from the analysis just outlined, the k-step transitional probabilities needed to estimate the probability that a certain predicted state will follow the present state. We thus have meteorological class prediction capabilities inherent within the normal data analysis routine.

The remaining problem is to predict the length of time interval that will be associated with the predicted meteorological class (and hence the variance level). The approach to this problem lies in determining whether the length of the time interval of successive meteorological classes is described by a dependent or independent random process. If the process is independent, then the best estimate for the length of the time interval of the predicted meteorological class is merely the mean value of the time interval frequency distribution for that class. If the process is found to be dependent on the preceding event, an algorithm similar to the one proposed for the variance estimates can be used. If the process is found to be dependent in a more complex manner, such as being dependent upon the length of the time interval for the previous 2 or 3 or n classes, then a probability model such as

$$LT_{i+1} = f_1 LT_i + f_2 LT_{i-1} + \cdots\cdots\cdots\cdots + f_n LT_{i-n} + r_{i+1}$$

---

[8] If the time series is found to be an independent random process, the $\mathrm{cov}[V_{T_i} V_{T_{i+1}}]$ would be zero.
The best advance estimate for the variance would then be the mean value of the distribution for the predicted meteorological class.

can be used,

where:

LT_{i+1} = length of the time interval to be predicted

LT_i = length of the time interval of the present meteorological class[9]

f = weighted factor of previous values of LT

r_{i+1} = the random element that must be taken into account

There are some distinct advantages to a program and an analysis such as the one proposed in this paper. The method is obviously extremely flexible, as mobile monitors can be moved wherever desired. The method can be adopted by almost any community with limited financial resources due to the variance-area relationship.[10] The analysis is basically simple and is Bayesian in nature. That is, new data are continually put into the system so that better and better estimates of predicted values can be realized. Most important, a mathematical statement of accuracy and confidence can be associated with each estimate of the upper bound concentration level for a specific pollutant.

Summary

A sampling strategy and methodology were proposed that enable an optimum size air monitoring surveillance network to be developed at a given cost. The estimates of pollutant concentrations that result from such a network are statistically accurate and valid to a predetermined level. The model develops two hypotheses into a prediction algorithm that enables advance estimates of the variance of pollutant concentration to be made. Knowing this, one is able to allocate the correct number of samples to best estimate pollution concentration levels with predetermined accuracy and confidence.

The mobilized monitoring approach taken by this strategy is probably the method that first should have been implemented. Unfortunately, the fast response time and portability features that were necessary for such an approach were not available when ambient air monitoring was first started. These requirements are being met by the new instrumentation available now, and planned for the future. If the mobilized approach by any methodology is deemed the proper way to go, we should not let tradition stand in the way.

---

[9] Note that LT_i must be estimated also, as we are dealing with a real time situation. That is, we don't really know how long LT_i will actually last until it ends. We can then get a revised estimate for LT_{i+1}.

[10] It should be noted that as a result of the analysis, the variance-area relationship becomes a family of curves—one curve for each classification. Thus the final graph depicted in Figure 1 will contain a family of curves.

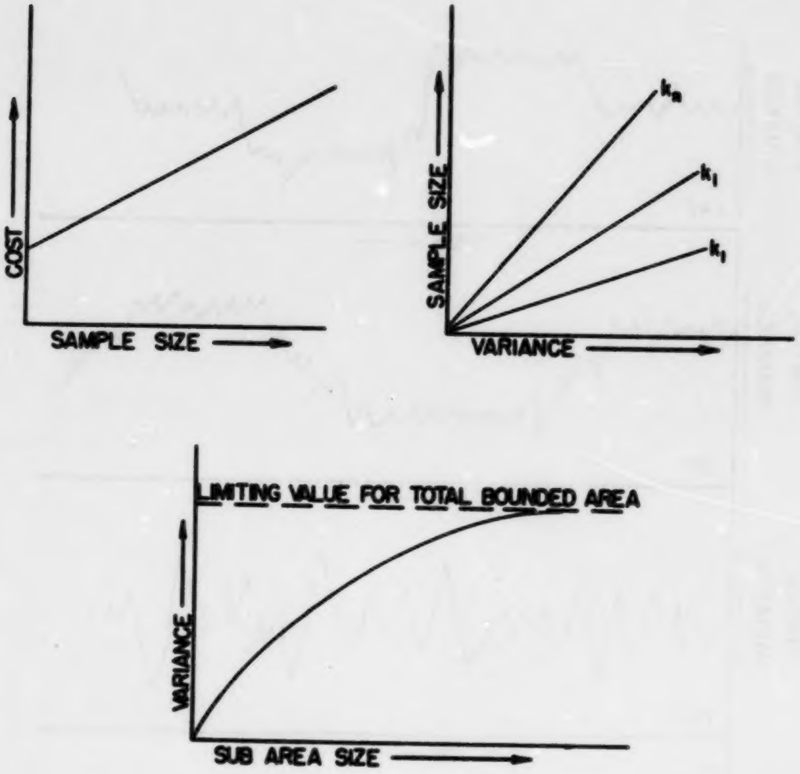


Figure 14-1. Use of variance-area relationship in sampling strategies.

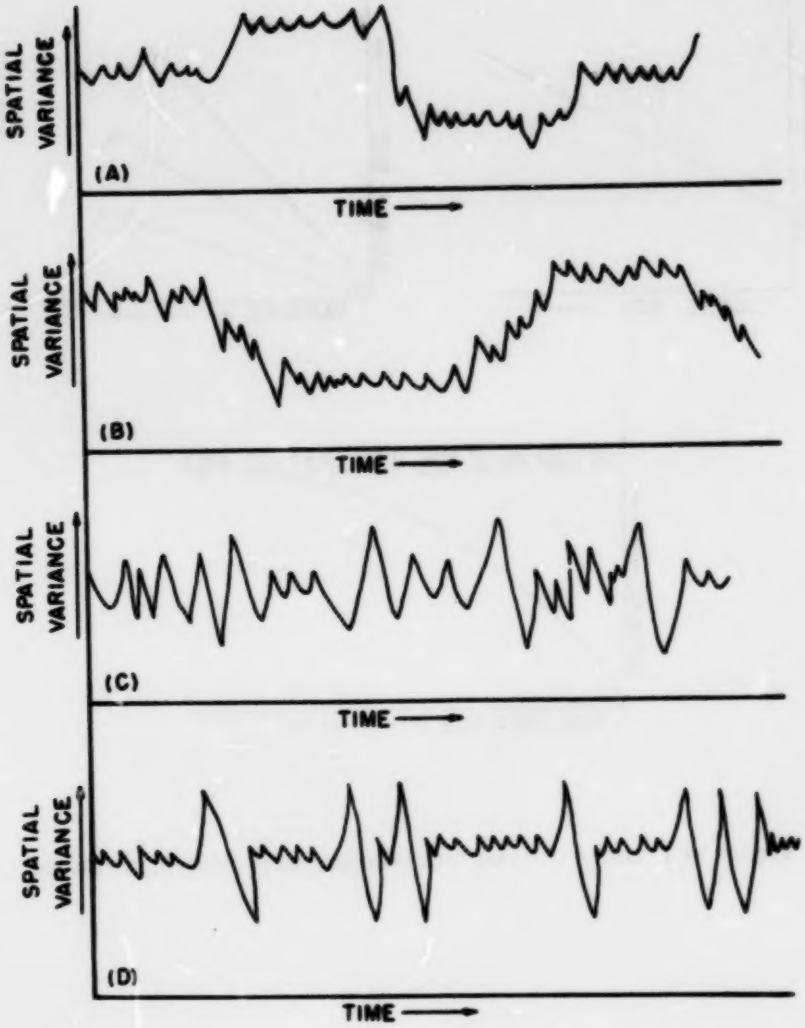


Figure 14-2. Time series representation of spatial variance.

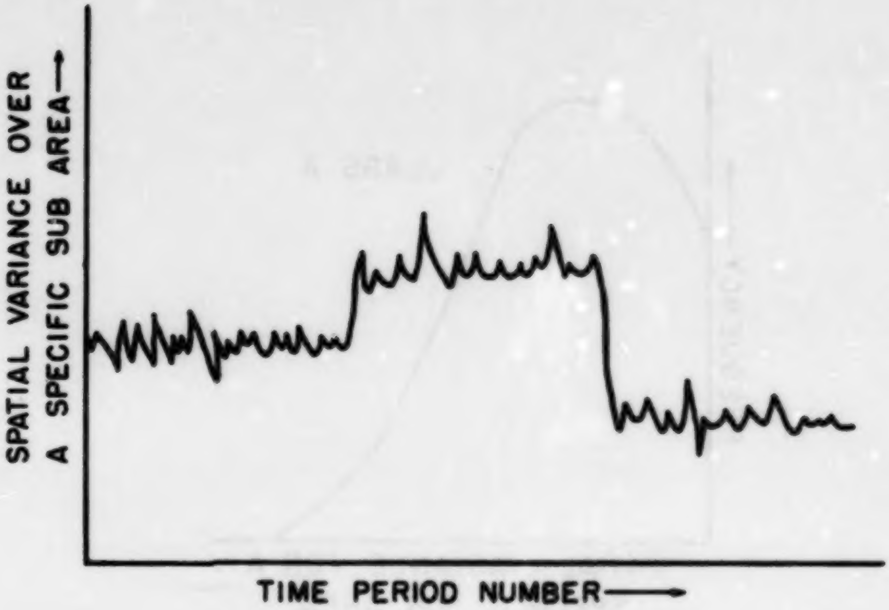


Figure 14-3. Spatial variance as a function of time.

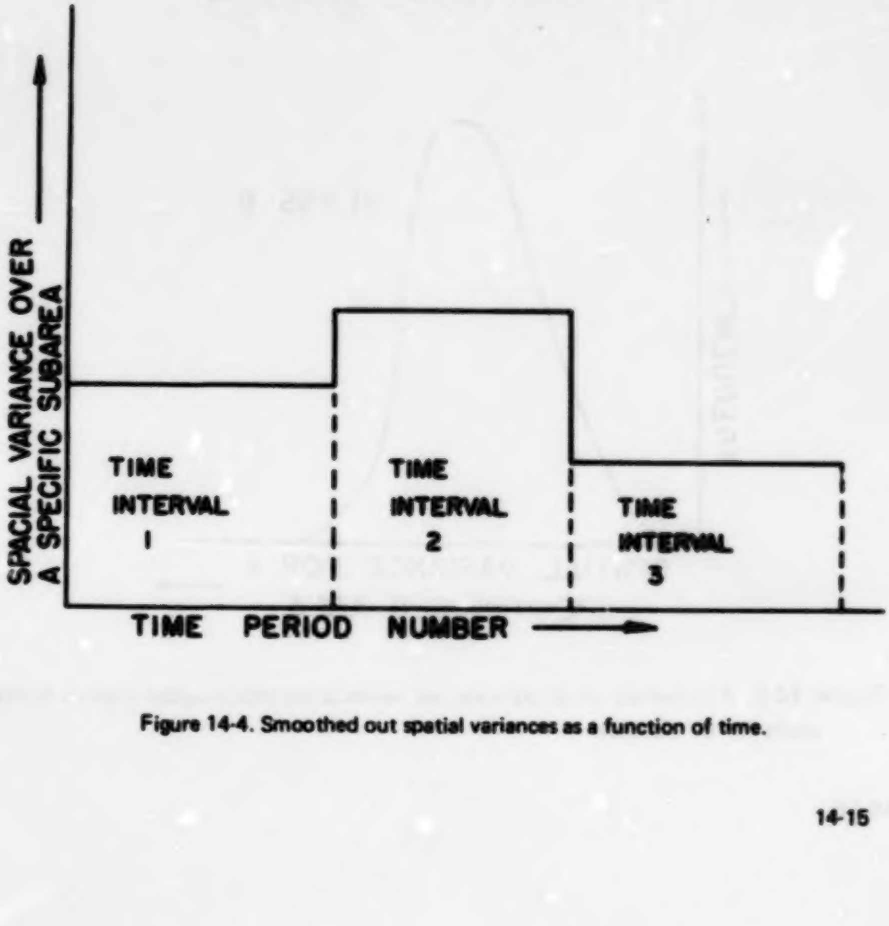


Figure 14-4. Smoothed out spatial variances as a function of time.

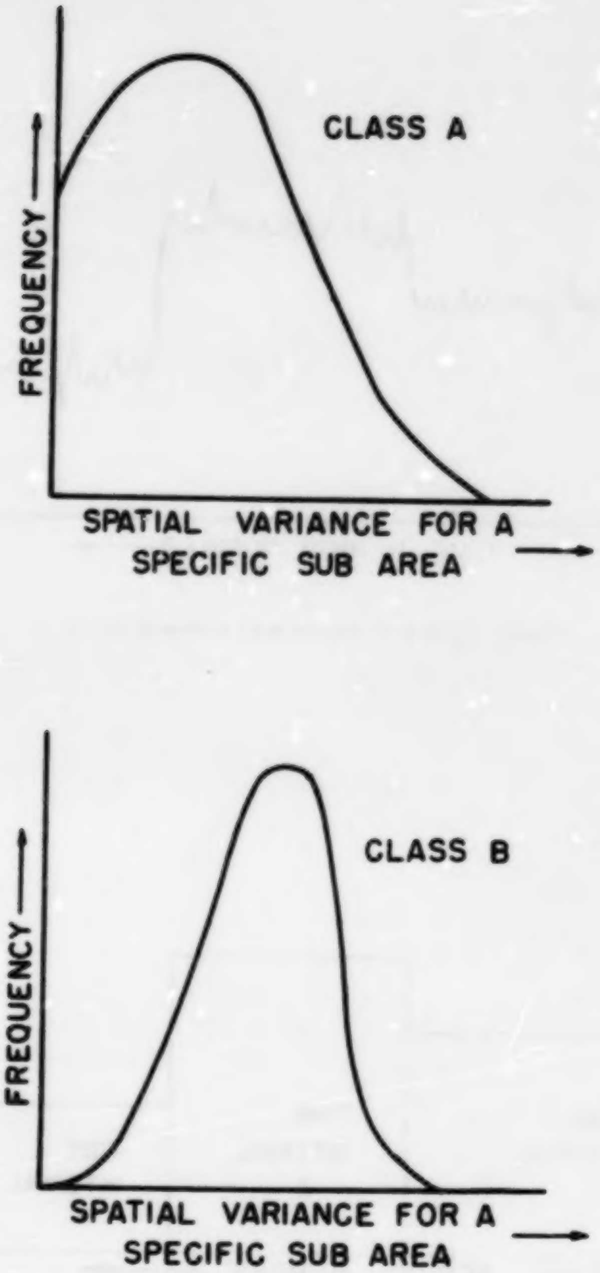


Figure 14-5. Frequency distributions for various meteorological classes (similar plots for all classes).

REFERENCES

Aitchison, J. and Brown, J.A.C., 1957: *The Lognormal Distribution.* Cambridge University Press.

Anon., 1971: Requirements for Preparation, Adoption, and Submittal of Implementation Plans. *Federal Register. 36:* No. 158.

Box, G.E.P., and Jenkins, G.M., 1971: *Time Series Analysis Forecasting and Control.* Holden-Day.

Cochran, W. G., 1963: *Sampling Techniques.* John Wiley.

Corn, M., 1970: Measurement of Air Pollution Dosage to Human Receptors in the Community. *Environmental Research. 3:* 218-233.

Davenport, W. B., 1970: *Probability and Random Processes.* McGraw-Hill.

E.P.A., 1971: Guidelines: Air Quality Surveillance Networks. United States Environmental Protection Agency, O.A.P., AP-98.

Larsen, R. I.,1971: A Mathematical Model for Relating Air Quality Measurements to Air Quality Standards. Environmental Protection Agency, O.A.P. Publication No. AP-89.

Pasquill, F., 1962: *Atmospheric Diffusion.* D. Van Nostrand Co. Ltd.

Stalker, W. W., Dickerson, R. C., and Kramer, G. D., 1962: Sampling Station and Time Requirements for Urban Air Pollution Surveys. *J. Air Pollution Control Association. 12:* 361-375.,

United States Atomic Energy Commission, 1968: Meteorology and Atomic Energy, 1968. A.E.C./Division of Technical Information.

DISCUSSION

Court: A third hypothesis that you should consider is that the variance, your spatial variance, will decrease as the sampling time increases. If you take samples for an entire year over your area, the difference betweeh the sampling sites will be much less than if you read them minute by minute. This all comes back then to the question of what you are trying to measure. If all that is wanted is the total concentration over an area, then your procedure may be valid. If we want to know where the hot spots are so we can track them down to a certain emitter, then we want to keep the stationary sampling sites. Furthermore, if we have stationary sites we can use the correlation between stations (the covariance) to look at the measure of the variability of the concentrations over area. A covariance between mobile stations would be meaningless whereas the covariance between fixed stations will indicate just how spotty the pattern is.

Visalli: Yes, I think you're absolutely correct. I would like to comment on one other thing, that is the concept of the biological half-life. If indeed the primary purpose of monitoring air is to protect public health, then the time period of analysis should be something that is going to be effective, that is, comparable to the biological half-life. In the case of SO_2 the biological half-life happens to be on the order of several minutes—I think 2-4 minutes. Well, if we're trying to get an estimate of what the air quality is in an area, I don't think that there is a way that we can do this by placing a monitor at a fixed point. We've got to have at least several monitors, in one area, whether fixed or not, to get an idea of what the probable concentration is for the entire area; and then see if we get meaningful cause and effect relationships. However I do agree with your point.

Marcus: One thing I would just like to point out is that as far as getting correlations between moving monitors, this seems to be a little bit analogous to the association between the Eulerian and Lagrangian approaches to correlation turbulence theory.

Dave Spiegler: One of the problems that National Weather Service has (I'm not with the National Weather Service) is to determine the optimum spacing of stations for an analysis which is similar to the problem you're talking about. What they do is have an objective analysis of an area and the spacing depends on the scale and how big the grid is. Here again it depends on whether, as Dr. Court said, you want to analyze over a period of an hour, or 6 hours, or a day, or what. The space scale is related to the time scale. The longer the period of time that you want to analyze, the fewer the number of stations necessary to describe the analysis over the area; i.e., the shorter wavelength features are less important as the time period increases. An objective analysis procedure has initial guess values on a grid of points and uses the station observations near those points to adjust the initial guess. I think that something like this might be useful in the work you are doing.

I. Enger: Just a small comment about the role of sampling, based upon just a very preliminary result that we came across a few weeks ago. Apparently a very small number of samples can approximate the distribution of an annual average very closely. What we did was for each 10 points we took a random sample of 100 and then a random sample of 10; and the means for two samples came very close. Not that I'm trying to say something statistical about it, just that it is an indication that a very small number of samples can estimate a distribution quite effectively. I think that's a point favorable to mobile sampling, provided you have the sampler returned to the same place. On a randomized scheme, 100 or 200 times a year you can get an annual average for a whole variety of places with only one instrument.

Wanta: I suppose that with the passage of time one will become acquainted with the characteristics of each of the sampling sites in the sense that one does with a single site. He learns what the near and more distant sources are.

15. THE EFFECT ON ROLLBACK MODELS DUE TO DISTRIBUTION OF POLLUTANT CONCENTRATIONS

YUJI HORIE AND JOHN OVERTON

Department of Environmental Sciences and Engineering
School of Public Health
University of North Carolina
Chapel Hill, North Carolina

Introduction

One of the important uses of air quality data is determination of the emission reduction required to achieve desired air quality. So-called "rollback equations" are often employed to compute a reduction in emissions. Basic questions have been raised on the application of rollback equations to emission standards for automobiles, and the effect of the distributions of present, future, and desired concentrations on the applicability of these equations.

The rollback equations that have been proposed are expressed as

$$R_L = \frac{g_L X_p - X_d}{g_L X_p - B} \tag{1}$$

$$R_J = \frac{g_J (X_p - B) - (X_d - B)}{g_J (X_p - B)} \tag{2}$$

where

R_L = reduction ratio (according to Larsen (1969))
R_J = reduction ratio (according to Jensen (1971))
X_p = present air quality
X_d = desired air quality
B = background concentration
g_L, g_J = growth factors over the period from the present to the goal year in which the desired air quality will be achieved.

The reason for the differences between these two equations is not being explored. The differences, however, are negligible when the background concentration, B, is much smaller than the present air quality, X_p. Both equations, for most practical purposes, can be reduced to

$$R \cong R_L \cong R_J \cong \frac{gX_p - X_d}{gX_p} \tag{3}$$

where $g \cong g_L \cong g_J$.

Once the reduction ratio, R, is determined, the permissible emissions can be calculated by using

$$e_d = e_f(1 - R) \tag{4}$$

where

e_d = future desired emission per unit source

e_f = future emission without controls per unit source.

If the future emission without controls, e_f, can be assumed to be the same as for the present (this is the case for automotive emissions), then Equation 4 can be written as

$$e_d = e_p(1 - R) \tag{5}$$

where

e_p = present emission per unit source.

The major question in using a rollback equation is how to compute the growth factor, g (g_L or g_J). According to Equation 1 the growth factor appears to be defined by

$$g_L \equiv \frac{X_f}{X_p} \tag{6}$$

where

X_f = future air quality without controls.

According to Equation 2 the growth factor appears to be defined by

$$g_J \equiv \frac{X_f - B}{X_p - B} \tag{7}$$

Values of pollutant concentration are described by both their averaging times and percentiles. The national air quality standards, for example, are stated as "D milligrams per cubic meter—maximum t-hour concentrations not to be exceeded more than once per year" (Anon. (1971)). Therefore the values of X_p, X_d, and B in the rollback equations must have the same averaging times, t-hours, etc. In addition to this, the growth factor defined either by Equations 6 or 7 should be calculated using the values of X_f, X_p, and B that correspond to the same percentile at which the value of X_d is designated, i.e., "once per year."

The usual method for determining the growth factor (Ott et al. (1967); Larsen (1961)) is: First, future emissions are estimated by some method such as projection of past car registrations to the goal year. Second, future concentrations are estimated by applying an air pollution display model to the future emissions using presently available meteorological data. Third, growth factor is determined from the ratio of the estimated future concentration to the present air quality that is either estimated by the same air pollution model or actually measured.

A question arises as to whether the ratio of future concentration to present concentration remains the same at every percentile value. This question may be restated—does the linear relation between emissions and concentrations, which leads to the rollback equations, extend to the percentile values of present, future and desired concentrations. In order to explore the assumption of linearity, which is implicitly employed in the usual method, a simple proportional model is assumed. As will be seen, a proportional model does not in general imply linearity between emissions and concentrations.

The discussion and consequence of this model are developed in the next section. Using the assumption of linearity over percentile values, the effect of concentration distributions on the rollback equations is investigated in the section titled "Rollback equations for percentile values." Several numerical examples are given in the succeeding section in which imaginary cities have been constructed and empirical distributions of the "city's" present and future concentrations have been calculated; growth factor for each percentile is graphically displayed showing the dependence of the growth factor on percentile and emission growth pattern. The implications of these calculations are discussed in the last section.

Simple Proportional Model

The air quality concentrations at any location in a city is assumed to be given by (Appendix I)

$$X = B + eF \qquad (8)$$

where

e = emission per unit source

F = function of all relevant variables such as weather factors and source distribution.

Even though B has a distribution, its value (even at the relevant high percentiles) is assumed much smaller than $e\,F$ and may be neglected. We assume it to be constant and independent of city growth.

In Equation 8, F is assumed not to be a function of e; thus the distribution of F determines the distribution of X. Future concentration at the α-th percentile, without emission controls, is

$$X_{f\alpha} = B + e_f F_{f\alpha} \tag{9}$$

$X_{f\alpha}$ depends, in part, on the future source distribution. This source distribution includes the effect of any city planning and regulatory practices. The desired future concentrations due to an emission reduction at the same percentile is

$$X_{d\alpha} = B + e_d F_{d\alpha} \tag{10}$$

Since F is independent of emission per unit source, we have

$$F_{d\alpha} = F_{f\alpha} \tag{11}$$

Using Equations 5 and 11, Equations 9 and 10 can be transformed to

$$X_{f\alpha} - B = e_p F_{f\alpha} \tag{12}$$

$$X_{d\alpha} - B = (1-R)\, e_p F_{f\alpha} \tag{13}$$

Thus we have

$$1 - R = \frac{X_{d\alpha} - B}{X_{f\alpha} - B} \tag{14}$$

The reduction ratio given by the above equation is independent of percentile, α. There is, however, no established means to estimate the percentile value of future concentration, $X_{f\alpha}$.

A growth factor, G_α, can be defined as

$$G_\alpha \equiv \frac{X_{f\alpha} - B}{X_{p\alpha} - B} \tag{15}$$

This growth factor is a measure of the growth of concentration value at different percentiles due only to the sources that can be controlled and thus excludes the background which is assumed to be independent of any growth. Substitution of the growth factor into Equation 14 yields

$$1 - R = \frac{X_{d\alpha} - B}{G_\alpha (X_{p\alpha} - B)} \tag{16}$$

Now the reduction ratio is expressed in terms of known variables except the growth factor.

Since $(X_{p\alpha} - B) = e_p F_{p\alpha}$ and $(x_{f\alpha} - B) = e_p F_{f\alpha}$, the growth factor can be written as (Appendix I)

$$G_\alpha = \frac{F_{f\alpha}}{F_{p\alpha}} \tag{17}$$

There is no reason to assume that the ratio, $(F_{f\alpha}/F_{p\alpha})$, is independent of percentile since both $F_{p\alpha}$ and $F_{f\alpha}$ are dependent on the emission inventories. The spatial distribution of emission sources changes as the city grows. Consequently, the distributions of $F_{p\alpha}$ and $F_{f\alpha}$ may be significantly different. In this sense, the proportional model does not necessarily imply a linear relationship between emissions and concentrations.

Rollback Equations for Percentile Values

This section is concerned with the dependence of the rollback equations on percentile and in particular the relation of these equations to the proportional model of the previous section.

For simplicity the following assumptions are made: (a) The growth factor is independent of percentile, i.e., $G = (F_{f\alpha}/F_{p\alpha}) =$ constant. (b) The rollback equations are valid for some percentile, say the 50-th. Then, the 50-th percentile reduction ratios in terms of any other percentile concentration values can be expressed as (Appendix II)

$$R_L = \frac{g_L X_{p\alpha} - X_{d\alpha} + \beta_\alpha (G - g_L)}{g_L X_{p\alpha} - B + \beta_\alpha (G - g_L)} \neq \frac{g_L X_{p\alpha} - X_{d\alpha}}{g_L X_{p\alpha} - B} \qquad (18)$$

and

$$R_J = \frac{g_J (X_{p\alpha} - B) - (X_{d\alpha} - B) + \beta_\alpha (G - g_J)}{g_J (X_{p\alpha} - B) + \beta_\alpha (G - g_J)} \qquad (19)$$

$$\neq \frac{g_J (X_{p\alpha} - B) - (X_{d\alpha} - B)}{G_J (X_{p\alpha} - B)}$$

where

$$\beta_\alpha = (X_{p\alpha} - X_{p50}) / e_p \qquad (20)$$

One can see that the form of Equations 18 and 19 is not invariant with a change in percentile α nor are R_L and R_J independent of α unless the growth factors, g_L and g_J are properly defined. A "natural" definition of g_L and g_J is

$$g_L \equiv \frac{F_{f\alpha}}{F_{p\alpha}}, \text{ and } g_J \equiv \frac{F_{f\alpha}}{F_{p\alpha}} \qquad (21)$$

In this case Equations 18 and 19 reduce to the same form as the original rollback equations, Equations 1 and 2.

In a more general case the growth factor, G, will be a function of percentile. In this situation the reduction ratio, R_j, defined in Equation 2 is the same as R in Equation 16 when the growth factor, g_J, is defined as in Equation 21, while the reduction ratio, R_L, is different from R even when the definition, Equation 21, is used.

Numerical Examples

A simple diffusion model (Hanna (1971)) is used to simulate the distribution of concentrations due to several source density configurations. In this model, as used, concentrations are a function only of the source density and wind velocity; i.e.,

$$X = \frac{c}{u}\left\{\rho_o + \sum_{i=1}^{N} \rho_i \left[(2i+1)^{.25} - (2i-1)^{.25}\right]\right\} \tag{22}$$

where

u = wind speed

ρ_i = density of sources in the $i^{\underline{th}}$ grid upwind from the central block "o"

c = constant

N = the number of upwind grid blocks included in the sum.

The source density, ρ, has been calculated using the sum of several Gaussian functions with the same parameters, each centered at different positions. This gives some structure to the "city." Future "cities" differ from present "cities" by the addition of a Gaussian function to the present city's source density configuration. The statistical distributions of concentrations are determined for each city by using Equation 22 with 2000 "wind" velocity vectors taken from a normal population; i.e., the X and Y components of wind velocity were computed independently by the use of a normal random number generator. From this pair of values the wind speed and direction were calculated. Concentrations were determined, using Equation 22, from 2000 pairs of wind speed and direction. Using the 2000 values of concentrations, empirical distributions were formed. Then the growth factors, G_α, were determined by the use of Equation 15, with zero background concentration, and the empirical distributions.

Three different source density configurations as shown in Figure 1 are used in the numerical examples. The present city's source densities (first source density configuration) are given by the sum of two Gaussian functions. The center of the second function is located at 5.0 kilometers east of that of the first. The receptor is at the center of the first function, the origin. The second configuration is the present city's configuration plus another Gaussian centered at 7.1 kilometers northeast of the origin. In the third configuration, the growth is represented by another Gaussian centered at 5.0 kilometers north of the origin.

The growth factors, G_{α}, have been calculated and plotted, in Figure 2 through 6, as a function of percentile for several normal velocity distributions. Of the 2000 values of growth factor, only seventy-five evenly spaced (along percentile axis) values have been plotted. On each figure are two sets of growth factors corresponding to the two different emission growth patterns. The solid circles give the growth factor when the future city is the second configuration, while the closed circles give the growth factor when the future city is the third configuration as described above. The wind velocity distribution is indicated on each diagram by $U_x \sim N$ (mean wind in x-direction, variance of x-wind component) and $U_y \sim N$ (mean wind in y-direction, variance of y-wind component). The growth factor diagram for the case of no preference in wind direction is shown in Figure 2. Beginning with a southerly dominant wind in Figure 3, the dominant wind direction keeps rotating clockwise by 90° for each of the remaining growth factor diagrams shown in Figures 4 through 6. Thus the effect of differing wind patterns relative to the cities' source configurations can be observed. The variances have been held constant, 25 and 100, respectively, for the wind components in East-West and North-South directions.

The growth factors appear to be more dependent on emission growth pattern than on concentration distribution. Although the two future cities have the same amount of increase in emissions, the growth factors of future city 2 (third configuration) are appreciably greater in every wind pattern and at every percentile than those of future city 1 (second configuration). The reason probably is that the third configuration (ρ ($\underline{r}$ = 0) = 1.90) yields a higher source density around the receptor point than the second configuration (ρ ($\underline{r}$ = 0) = 1.65) does, and that pollutant concentrations are influenced in a greater extent by nearby sources than remote sources.

The dependence of the growth factor on percentile is fairly sensitive on the source configuration relative to the wind pattern. As the dominant wind direction rotates, the slope of the trends changes. Strong dependence on percentile occurs when the dominant wind blows from the North or South (see Figures 3 and 5 respectively). Weak or no dependence on percentile occurs when the dominant wind blows from the East or West. When wind does not have a preferred direction the growth factors show a mild dependence on percentile.

It is difficult to determine how realistic the model simulations are. The wind roses were constructed from the wind speeds and wind directions generated by the method mentioned above. The simulated wind frequency distributions are not much different in nature from those observed at CAMP cities.

Discussion and Conclusions

The effect of statistical distribution of pollutant concentrations on rollback equations has been investigated by using the concept of the proportional model as well as by using numerical examples. The form of the rollback equations, in general, changes with percentiles of pollutant concentration when incorrect growth factors are used. This conclusion derives from the percentile forms of the rollback equations that has been obtained for the case of a constant growth factor by using the proportional model.

Most cities may not be expected to grow in any simple fashion. To see the dependency of the growth factor on percentiles and emission growth patterns, the growth factors at different percentiles of an imaginary city were calculated for several emission growth patterns and for several different wind velocity distributions. The percentiles were constructed from 2000 concentration values that were calculated using a simple model and normal random wind velocity vectors as meteorological input. The results are shown in Figures 2 through 6.

In all cases the figures show that the growth factor of future city 2 (open circles) are greater than those of future city 1 (solid circles) although the amount of emission growth is the same for the two future cities. The reason is that emission growth around the receptor point is larger in the 2nd future city than in the first, and that the receptor concentrations are affected in a greater extent by nearby sources than remote sources. Here, the source densities at the origin where the receptor is located are 1.45, 1.65 and 1.90, respectively, for present city, future city 1, and future city 2 (Fig. 1). The "emission growth factors" at the receptor point, therefore, are 1.14 and 1.31, respectively, for future cities 1 and 2. This indicates that not only the amount of emission growth but also the emission growth pattern is important to correctly estimate the growth factor. This in turn suggests that redistribution of emission sources through city planning can be an effective measure to improve the air quality at dirty spots in a city.

The growth factor may increase or decrease with percentile depending upon the emission growth pattern and wind pattern. When there is no preference in wind direction (Fig. 2), the growth factors gradually increase with percentile. The reason is that a weak wind, which results in higher concentrations, tends to magnify the effect of emission growth on concentrations. When a dominant wind blows from the East or West (Figs. 4 and 6), the growth factors become less dependent on percentile. Strong dependence on percentile occurs when the dominant wind blows from the North or South. When the dominant wind blows from the South (Fig. 3), the growth factors increase with percentile. For a northerly dominant wind the growth factors decrease with percentile (Fig. 5). This can be explained as follows:

(a) Lower percentile concentrations in Figure 3 result from a strong southerly wind that blows over the southern part of the city where the emission growth is lower than the other part of a city. Thus, the concentration growth factors at lower percentiles are smaller than the "emission growth factors" at the receptor.

(b) Higher percentile concentrations in Figure 3 result from a weak northerly wind that carries pollutant from the northern part of the city where the emission growth is higher than the other part. Thus, the concentration growth factors at higher percentiles are greater than the "emission growth factor" at the receptor.

(c) As a result of (a) and (b), the growth factors in Figure 3 increase rapidly with percentile.

(d) Lower percentile concentrations of future city 2 result from a strong northerly wind that carries pollutant from the high emission growth northern part to the receptor. Thus, the concentration growth factors at lower percentiles are much greater than the "emission growth factor" at the receptor (=1.31). On the other hand, higher percentile concentrations of future city 2 result from a weak southerly wind that blows over the low emission growth southern part of a city. Since a weak wind tends to magnify the effect of emission growth on concentrations as mentioned before, the concentration growth factors at higher percentiles are about the same or a little higher than the "emission growth factor" at the receptor. As a result of this, the growth factors of future city 2 decrease sharply with percentile as seen from Figure 5.

(e) The growth factors of future city 1 decrease with percentile by a similar reason to the above. However, the downward trend is much milder than that of future city 2 because the center of the emission growth is located at the northeast of the receptor instead of the North, and is more distant from the receptor point than that of future city 2.

From the preceding discussion the following qualitative statement can be made as to percentile dependence of growth factors. When there is no preference in wind direction and speed, growth factors tend to increase with percentile unless source density configuration is point symmetric and the receptor is located at the symmetric point. When there is a dominant wind direction from which wind blows strong and frequently, growth factors at downwind receptors from the emission growth center decrease with percentile and those at upwind receptors increase with percentile.

The effects of concentration distribution and emission growth pattern on growth factors have been discussed mainly because the growth factors are used extensively in the literature (Larsen (1961) (1969); Jensen (1971); Ott et al. (1967)). The real concern, however, is the effects of concentration distribution and growth pattern on the reduction ratio that is given either by Equations 1 or 2 or 16. The sensitivity of the reduction ratio, R, on the growth factor, G, can *be checked by expanding R in a Taylor series about the correct value of G,* G_o, and in terms of the deviation of G from the correct value, $\delta = G - G_0$.

$$R(G) \simeq R(G_o) + \delta \frac{\delta R}{\delta G}\bigg|_{G_o} \tag{23}$$

Setting B = 0 in Equations 1, 2, and 16 we can obtain

$$\frac{dR}{dG}\bigg|_{G_o} = \frac{1 - R(G_o)}{G_o} \tag{24}$$

Thus we can write the reduction ratio as

$$R(G) \simeq R(G_o) + \frac{\delta(1 - R[G_o])}{G_o} \tag{25}$$

Suppose that one made a mistake in estimating future growth pattern and ignored the effect of concentration distribution on growth factors. Thus, from Figure 3 the difference between the growth factors at the 50-th percentile of future city 1 and the 99-th percentile of future city 2 is about –0.25. Taking R $(G_0) = 0.90$, then from Equation 25, $R(G) \cong 0.87$. This appears to be a negligible effect. Real importance, however, is as to how much the total amount of pollutants will remain when emissions are reduced according to the reduction ratio. The relative difference in the amounts of remaining pollutants according to the correct and incorrect reduction ratios is given by

$$\frac{\left\{(1-R[G_o]) - (1-R[G])\right\} e_p \int_{\underline{r}} \rho_f (dr')^2}{(1-R[G_o]) e_p \int_{\underline{r}'} \rho_f \,(dr')^2} \tag{26}$$

$$= \frac{R(G) - R(G_o)}{1 - R(G_o)}$$

Substituting the values into Equation 26, the example yields about 30% difference in the amounts of remaining pollutants. This can not be a negligible effect.

The results of this work are not conclusive as to the extent of the effects of concentration distribution and emission growth pattern on calculating reduction ratios. It is, however, obvious that a correct reduction ratio can not be obtained from emission growth only. Consideration on emission growth patterns and concentration distributions should be included in rollback calculations.

$$\rho(\underline{r}) = \sum_{n=1}^{N} \rho_n(\underline{r}-\underline{R}_n)$$

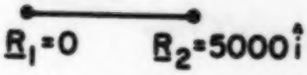
$$\rho_n(\underline{r}-\underline{R}_n) = \mathrm{EXP}\left[-10^{-8}\pi(\underline{r}-\underline{R}_n)^2\right]$$

Present City:

$N=2$

$\rho=\rho_1+\rho_2$

$\rho(\underline{r}=0)=1.45$

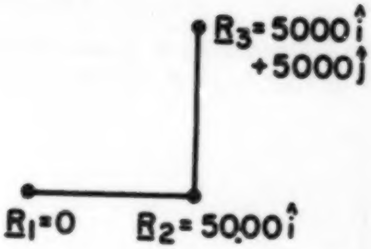

$\underline{R}_1=0$
$\underline{R}_2=5000\hat{i}$

Future City 1:

$N=3$

$\rho=\rho_1+\rho_2+\rho_3$

$\rho(\underline{r}=0)=1.65$


$\underline{R}_3=5000\hat{i}+5000\hat{j}$
$\underline{R}_1=0$
$\underline{R}_2=5000\hat{i}$

Future City 2:

$N=3$

$\rho=\rho_1+\rho_2+\rho_3$

$\rho(\underline{r}=0)=1.90$

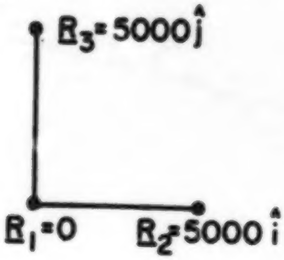

$\underline{R}_3=5000\hat{j}$
$\underline{R}_1=0$
$\underline{R}_2=5000\hat{i}$

Figure 15-1. Source configuration of imaginary city

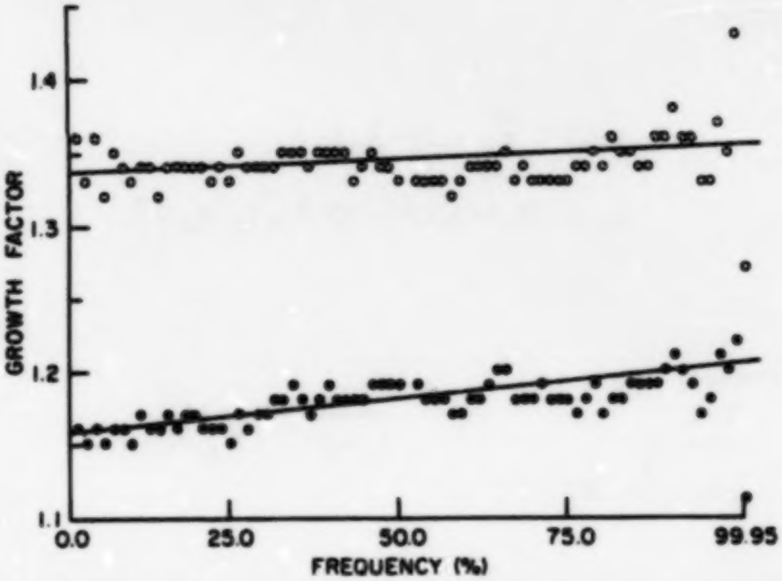


Figure 15-2. Cumulative frequency diagram for the growth factor for case I: $U_x \sim N(0,100)$, $U_y \sim N(0,100)$

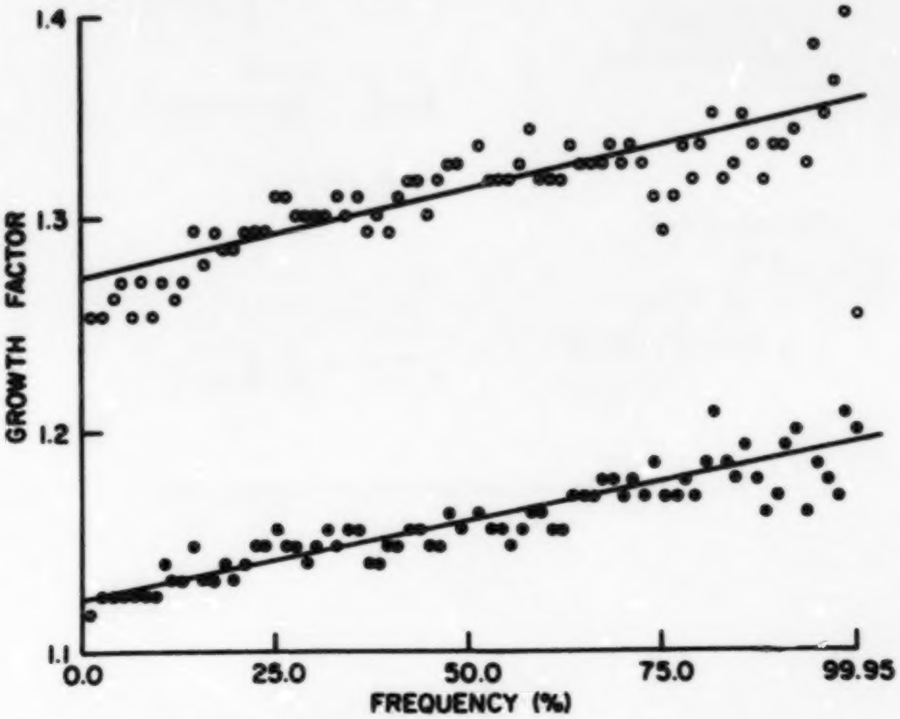


Figure 15-3. Cumulative frequency diagram for the growth factor for case II: $U_x \sim N(0,25)$, $U_y \sim N(5,100)$

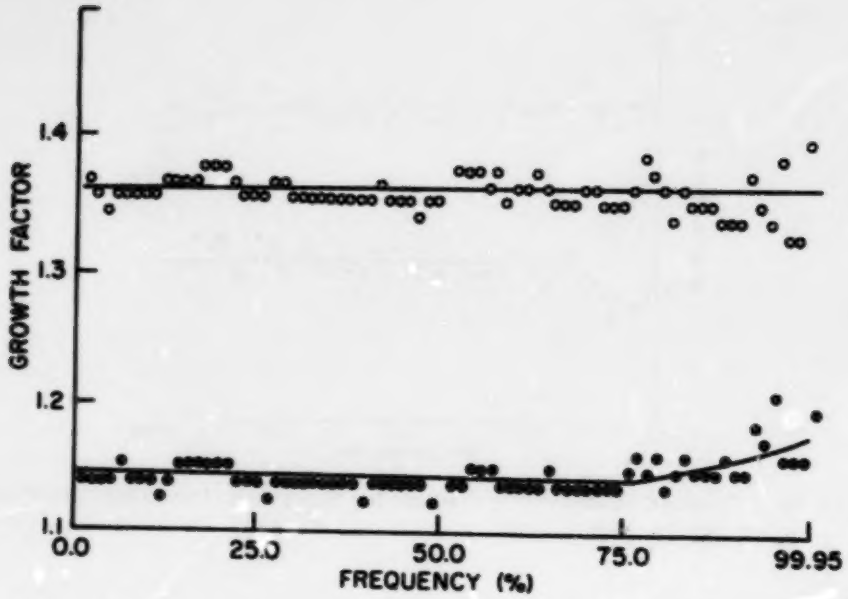


Figure 15-4. Cumulative frequency diagram for the growth factor for case III: $U_x \sim N(5,25)$, $U_y \sim N(0,100)$

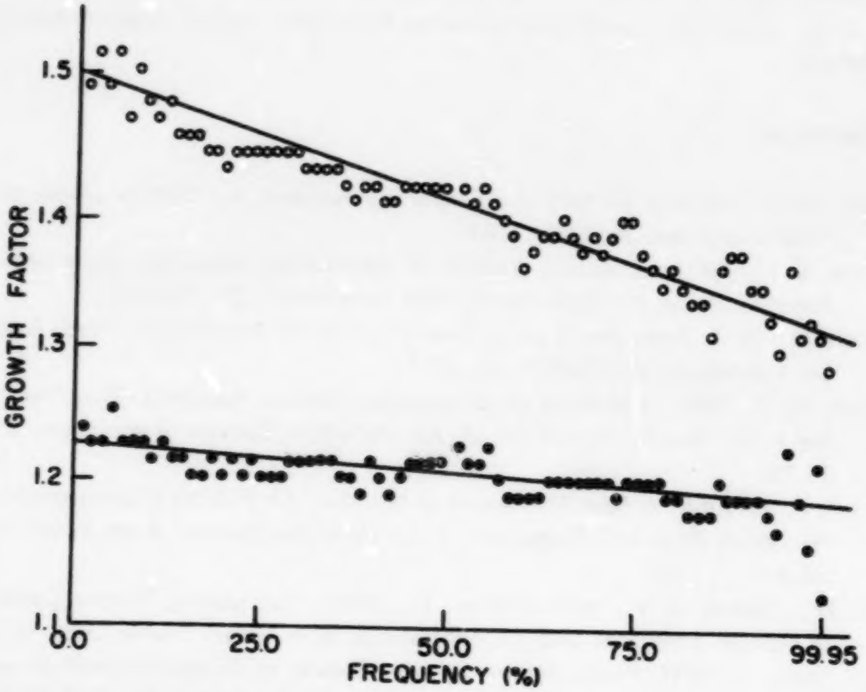


Figure 15-5. Cumulative frequency diagram for the growth factor for case IV: $U_x \sim N(0,25)$, $U_y \sim N(-5,100)$

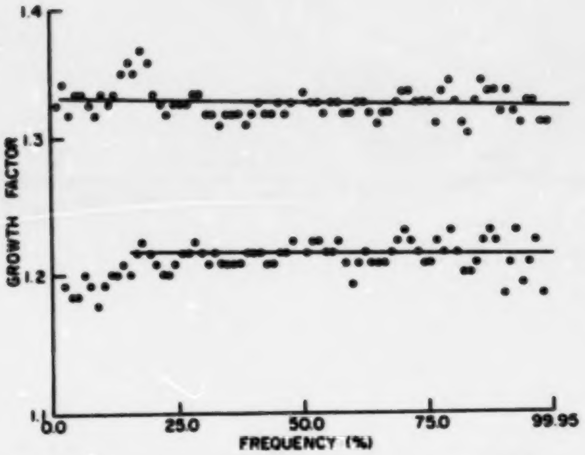


Figure 15-6. Cumulative frequency diagram for the growth factor for case V: $U_x \sim N(-5,25)$, $U_y \sim N(0,100)$

Acknowledgments

The authors wish to thank Mr. Richard Kamens for fruitful discussions and Professor Arthur C. Stern for his encouragement and helpful suggestions. This work was supported by the Environmental Protection Agency research project R-800901.

References

Anon, 1971: National Primary and Secondary Ambient Air Quality Standards. *Federal Register. 36:* 8187-8188.

Hanna, S. R., 1971: A Simple Method of Calculating Dispersion from Urban Area Sources. *J. Air Pollution Control Association. 21:* 774-777.

Jensen, D., 1971: From Air Quality Criteria to Control Regulations. Ford Motor Co. Publication No. 710303, pp. 67-74.

Larsen, R. I., 1961: A Method for Determining Source Reduction Required to Meet Air Quality Standards. *J. Air Pollution Control Association. 11:* 71-76.

Larsen, R. I., 1969: A New Mathematical Model of Air Pollutant Concentration Averaging Time and Frequency. *J. Air Pollution Control Association. 19:* 24-30.

Ott, W., Clarke, J. F., and Ozolins, G., 1967: Calculating Future Carbon Monoxide Emissions and Concentrations from Urban Traffic Data. U. S. Dept. of HEW, Public Health Service, Bureau of Disease Prevention and Environmental Control, Public Health Service Publication No. 999-AP-41.

APPENDIX I

Simple Multiple-Source Model

The concentrations at position $\underline{r}$ and time t due to all relevant physical factors $\underline{w}$ can be expressed for a single type of sources as

$$X_s(\underline{r},t,\underline{w}) = e\int_{t'}\int_{\underline{r}'}\rho\,(\underline{r}',t',\underline{w})\;A(\underline{r}-\underline{r}',\,t-t',\,\underline{w})(dr')^2dt'$$

where (I-1)

X_s = concentration due to one type of source
e = emission per unit source
ρ = density of emission sources
A = transfer function that relates sources at $\underline{r}'$, t' to concentration at $\underline{r}$, t.

The total concentration is given by the sum of the background concentration, B, and the concentration due to the one type of source, X_s. Defining a function F as (I-2)

$$F(\underline{r},t,\underline{w}) \equiv \int_{t'}\int_{\underline{r}'}\rho\,(\underline{r}',t',\underline{w})\;A(\underline{r}-\underline{r}',\,t-t',\,\underline{w})(dr')^2dt'$$

the total concentration can be written as

$X = B + e\,F$ (8)

Equation 8 indicates that the distribution of the total concentration, X, is determined by the distribution of F through the physical factors $\underline{w}$, position $\underline{r}$, and time t. Thus the a-th percentile value of X, X_a, is determined by the a-th percentile of F, F_a, where the background concentration, B, is assumed to be constant.

The growth factor defined by the ratio of $(X_{fa} - B)/(X_{pa} - B)$ can be expressed as

$$G_a \equiv \frac{F_{fa}}{F_{pa}}$$

$$= \frac{\int_{t'}\int_{\underline{r}'}\rho_f\,(\underline{r}',t',\underline{w}_f)\;A(\underline{r}-\underline{r}',\,t_f-r',\,t_f-t',\,\underline{w}_f)(dr')^2dt'}{\int_{t'}\int_{\underline{r}'}\rho_p\,(\underline{r}',t',\underline{w}_p)\;A(\underline{r}-\underline{r}',\,t_p-t',\,\underline{w}_p)(dr')^2\,dt'} \qquad \text{(I-3)}$$

where the integrals in Equation I-3 can be computed with variables, t and $\underline{w}$, that would give the a-th percentiles of F_p and F_f.

APPENDIX II

Derivation of the Percentile Form of the Rollback Equations

Assume that $G = (F_{f\alpha}/F_{p\alpha})$ is independent of percentile α and that the rollback equations are valid for the 50-th percentile, i.e.,

$$R_L = \frac{g_L X_{p50} - X_{d50}}{g_L X_{p50} - B} \qquad \text{(II-1)}$$

$$R_J = \frac{g_J (X_{p50} - B) - (X_{d50} - B)}{g_J (X_{p50} - B)} \qquad \text{(II-2)}$$

where

X_{p50} = the 50-th percentile value of X_p

X_{d50} = the 50-th percentile value of X_d.

The proportional model, in general, can be written as

$$X_i = B + e_i F_i \qquad \text{(II-3)}$$

where subscript i is used to generalize quantities for future, present and desired. From the preceding discussions we have

$$F_{f\alpha} = F_{d\alpha} = GF_{p\alpha} \qquad \text{(II-4)}$$

$$e_f = e_p$$

$$e_d = (1 - R)\, e_p$$

$$R = R_L \text{ or } R_J$$

The α-th and 50-th percentiles of the present and desired air qualities can be obtained, using R_L, from Equations II-3 and II-4 as

$$X_{p\alpha} = B + e_p F_{p\alpha} \qquad \text{(II-5)}$$

$$X_{d\alpha} = B + (1 - R_L)\, e_p G F_{p\alpha} \qquad \text{(II-6)}$$

$$X_{p50} = B + e_p F_{p50} \qquad \text{(II-7)}$$

$$X_{d50} = B + (1 - R_L)\, e_p G F_{p50} \qquad \text{(II-8)}$$

Any percentile value of F_p can be related to the 50-th percentile value of F_p, F_{p50}, by a constant, β_α.

$$F_{p\alpha} = F_{p50} + \beta_\alpha \tag{II-9}$$

Substitutions of Equations II-5 and II-9 into Equation II-7 and Equations II-6 and II-9 into Equation II-8 yield, respectively,

$$X_{p50} = X_{p\alpha} - e_p \beta_\alpha \tag{II-10}$$

and

$$X_{d50} = X_{d\alpha} - (1 - R_L) e_p G \beta_\alpha \tag{II-11}$$

Substitution of Equations II-10 and II-11 into Equation II-1 yields

$$R_L = \frac{g_L (X_{p\alpha} - e_p \beta_\alpha) - (X_{d\alpha} - [1 - R_L] e_p G \beta_\alpha)}{g_L (X_{p\alpha} - e_p \beta_\alpha) - B} \tag{II-12}$$

Solving for R_L we obtain

$$R_L = \frac{g_L X_{p\alpha} - X_{d\alpha} + \beta_\alpha (G - g_L)}{g_L X_{p\alpha} - B + \beta_\alpha (G - g_L)}$$

By similar steps Equation 19 can be derived from Equation II-2.

DISCUSSION

Larsen: Thank you, Dr. Horie, for a thorough look at this problem. We might consider one factor. Three rollback equations could be used according to the behavior of background concentrations. We have heard of two rollback equations, R_J which refers to Jensen's article and R_L which refers to Larsen's equation. Jensen's equation is the correct equation if background now remains the same as background later; a second rollback equation would be one in which the concentration might be doubling and the background would also be doubling. In other words, a second rollback equation would be one in which the background growth factor equaled the urban growth factor. The Larsen rollback equation is between these two, between a background growth factor of one and a background growth factor equal to the urban growth factor. For a situation involving no background growth, the Jensen equation should be used. If growth is intermediate the Larsen equation should be used. This intermediate situation might be experienced, for instance, with a northeast wind blowing from Boston to New York to Philadelphia to Baltimore to Washington. As these places grow together, the background grows together and background increases. The Larsen equation could be used for places growing together. In the middle of the great plains, not affected by background from other cities, the Jensen equation could be used.

Horie: Thank you very much. This is exactly so. We noticed this difference when we consider the growth for the background concentration.

Smith: I think one fact which ought be taken into account in this sort of calculation is the variation due to the modification of the weather by the pollutants themselves. If you reduce the concentration of such things as smoke or photochemical components, then you may very well change the statistics of the weather and hence, get a change in your reduction factor.

16. SYMPOSIUM PARTICIPANTS

Pat Adomitis
Dept. of ESE
School of Public Health
University of North Carolina
Chapel Hill, N. C. 27514

Gerald G. Akland
EPA – Div. of Atmospheric Surveillance
Research Triangle Park, N. C. 27711

James N. Arvesen
Dept. of Statistics
Math Science Bldg.
Purdue University
Lafayette, Indiana 47907

C. W. Ash
Dept. of Biostatistics
School of Public Health
University of North Carolina
Chapel Hill, N. C. 27514

R. E. Barlow
University of California
Berkeley, Ca. 94700

Joel Barnett
Div. of Air Pollution Control
C2-212 Cordell Hull Building
Nashville, Tennessee 37219

John J. Beauchamp
Oak Ridge National Laboratory
Statistics Department
P. O. Box Y
Bldg. 9704 -1
Oak Ridge, Tennessee 37830

Michael M. Benarie
IRCHA Research Center
Boite Postale 1
91 Vert-le-Petit
France

Bernard Bloom
Allegheny Cty. Air Poll. Control Board
301 39th St.
Pittsburg, Penn. 15201

John M. Bowman
501 N. 9th St.
Room 130
Safety, Health & Welfare Building
Richmond, Virginia 23219

Franklin Briese
Div. of Biometrics
Mail Container 2355
Univ. of Colorado Med. Center
Denver, Colorado 80220

Kenneth Calder
EPA – Div. of Meteorology
Research Triangle Park, N. C. 27711

Norman L. Canfield
19 Bayberry Ave.
Garden City, New York 11530

Charles R. Case
TRC
125 Silas Deane Highway
Wethersfield, Conn. 06109

C. Andrew Clayton
EPA
Research Triangle Park, N. C. 27711

David C. Collins
Technology Service Corp.
225 S. Monica Blvd.
Santa Monica, Ca. 90401

R. E. Cooper
Environmental Analysis and Planning
Savannah River Laboratory
Aiken, S. C. 29801

Arnold Court
California State University
Northridge, California

J. M. Craig
Southern Services Inc.
P. O. Box 2625
Birmingham, Alabama 35202

Alexander R. Craw
National Bureau of Standards
Washington, D. C. 20234

T. V. Crawford
Environmental Analysis and Planning
Savannah River Laboratory
Aiken, S. C. 29801

Loren W. Crow
2422 South Downing St.
Denver, Colorado 80210

E. James Dale
Dept. of ESE
School of Public Health
University of North Carolina
Chapel Hill, N. C. 27514

Eugene M. Darling, Jr.
DOT Transportation Systems Center
55 Broadway
Cambridge, Massachusetts 02147

William Delaware
Dept. of Envir. Conservation
Div. of Air Resources
50 Wolf Road
Albany, New York

John B. Edwards
Chrysler Corp.
CIMS 418-18-22
P. O. Box 1118
Detroit, Michigan 48231

Isodore Enger
Geomet, Inc.
50 Monroe St.
Rockville, Maryland 20850

James A. Fay
Room 3246
MIT
Cambridge, Massachusetts 02139

Robert Faoro
EPA
Research Triangle Park, N. C. 27711

Martin A. Ferman
Research Laboratories
Department E.V.
General Motors Technical Center
Warren, Michigan 48090

Doug Fox
EPA/NERC
Research Triangle Park, N. C. 27711

Neil H. Frank
EPA
Research Triangle Park, N. C. 27711

Joel Frockt
Department of Biostatistics
School of Public Health
University of North Carolina

A. S. Galbraith
EPA
Research Triangle Park, N. C. 27711

Frank A. Gifford
NOAA-ATDL
P. O. Box E
Oak Ridge, Tennessee 37830

Steve Goranson
Office of Statistical Services
EPA/NERC
Research Triangle Park, N. C. 27711

Rebecca A. Gray
EPA/NERC
Div. of Chemistry and Physics
Research Triangle Park, N. C. 27711

George W. Griffing
EPA/Div. of Meteorology
Research Triangle Park, N. C. 27711

Nathaniel Guttman
National Climatic Center
Federal Building
Asheville, North Carolina 28801

Howard R. Hammond
Baltimore Gas and Electric
Room 231
531 E. Madison St.
Baltimore, Maryland

David M. Hershfield
ARS -W Soils Building
Beltsville, Maryland 20705

C. Doyce Hester
Reynolds Metals Co.
P. O. Box 9177
Corpus Christi, Texas 78408

M. Eugene Hoffman
North Carolina State University
Raleigh, North Carolina 27607

George C. Holzworth
EPA/ Div. of Meteorology
Research Triangle Park, N. C. 27711

Yuji Horie
ESE, School of Public Health
University of North Carolina
Chapel Hill, N. C. 27514

William F. Hunt
EPA
Research Triangle Park, N. C. 27711

Peter Imrey
Department of Biostatistics
School of Public Health
University of North Carolina
Chapel Hill, N. C. 27514

John S. Irwin
University of North Carolina
Chapel Hill, N. C. 27514

Warren G. Johnson
EPA
Research Triangle Park, N. C. 27711

Howard C. Jones
Dept. of Environmental Conservation
Div. of Air Resources
50 Wolf Road
Albany, New York

James Jones
Dept. of Chemical Engineering
University of Kentucky
Lexington, Kentucky

Surendra Joshi
ESE, School of Public Health
University of North Carolina
Chapel Hill, N. C. 27514

R. H. Ketterer
General Electric
Schenectady, New York

K. R. Knoerr
308 Biological Sciences
Duke University
Durham, North Carolina 27706

Joseph B. Knox
Lawrence Livermore Laboratory
University of California
P. O. Box 808
Livermore, California 94550

Lewis Kontnik
ESE, School of Public Health
University of North Carolina
Chapel Hill, North Carolina

Stanley L. Kopczynski
EPA/NERC
Research Triangle Park, N. C. 27711

Lawrence L. Kupper
Dept. of Biostatistics,
School of Public Health
University of North Carolina
Chapel Hill, N. C.

Ralph Larsen
EPA/Met. Laboratory
Research Triangle Park, N. C. 27711

Russell F. Lee
EPA/OAP
Research Triangle Park, N. C. 27711

Stephen A. Lesh
ESE, School of Public Health
University of North Carolina
Chapel Hill, N. C. 27514

Helmut Lieth
Department of Botany
University of North Carolina
Chapel Hill, N. C. 27514

Jiumn W. Lin
Room 402,
Dept. of Environ. Control
320 N. Clark St.
Chicago, Illinois 60610

James W. Lingle
Industrial Bio-Test Labs Inc.
1810 Frontage Rd.
Northbrook, Illinois 60002

Gene R. Lowrimore
EPA
Research Triangle Park, N. C. 27711

David A. Lynn
33 Cogswell Ave.
Cambridge, Massachusetts 02140

George Manire
EPA
Research Triangle Park, N. C. 27711

Allan H. Marcus
Dept. of Mathematical Sciences
Johns Hopkins University
Baltimore, Maryland 21218

David McLeod
EPA
Research Triangle Park, N. C. 27711

Thomas E. McMullen
EPA
Research Triangle Park, N. C. 27711

Donald McNeil
Dept. of Statistics
Princeton University
Princeton, New Jersey 08540

David McNelis
Dept. of ESE
University of North Carolina
Chapel Hill, N. C. 27514

W. S. Meisel
Technology Service Corp.
225 Santa Monica Blvd.
Santa Monica, Ca. 90401

Edwin L. Meyer
EPA/OMP
Research Triangle Park, N. C. 27711

Mark Murray
Dept. of Biostatistics
School of Public Health
University of North Carolina
Chapel Hill, N. C. 27514

Carl Nelson
Research Triangle Institute
Box 12194
Research Triangle Park, N. C. 27711

Harold E. Neustadter
NASA-LERC
21000 Brookpack Road
Cleveland, Ohio 44135

Everett C. Nickerson
EPA
Research Triangle Park, N. C. 27711

Lawrence Niemeyer
EPA/Div. of Meteorology
Research Triangle Park, N. C. 27711

John Overton
ESE, School of Public Health
University of North Carolina
Chapel Hill, N. C. 27514

Don Pack
NOAA
8060 13th Street
Silver Springs, Maryland 20910

John J. Paulus
Westinghouse Electric Corp.
P. O. Box 9533
Raleigh, North Carolina

Robert Papetti
EPA
Waterside Mall
4th and M St. S.W.
Washington, D. C.

M. M. Pendergrast
Envir. Analysis and Planning
Savannah River Laboratory
Aiken, S. C. 29801

Bier Peters
North Carolina State University
Raleigh, N. C. 27607

James Peterson
EPA/Div. of Meteorology
Research Triangle Park, N. C. 27711

John M. Pierrard
Petroleum Laboratory
E. I. DuPont de Nemours
Wilmington, Delaware 19898

Richard I. Pollack
University of California
Lawrence Livermore Laboratory
Livermore, California

Charles Proctor
Statistics Department
North Carolina State University
Raleigh, North Carolina 27607

William J. Raynor
Dept. of Biostatistics
School of Public Health
University of North Carolina
Chapel Hill, N. C. 27514

John H. Reynolds
2615 Selwyn Ave.
Charlotte, N. C. 28209

Donald M. Rote
Argonne National Laboratory
9700 S. Cass Ave.
Argonne, Illinois 60439

R. Ruff
EPA/Div. of Meteorology
Research Triangle Park, N. C. 27711

J. S. Rustagi
Division of Statistics
The Ohio State University
Columbus, Ohio 43210

Bernard E. Saltzman
Kettering Laboratory
Eden and Bethesda Ave.
Cincinnati, Ohio 45219

Don Thomas
Air Management Branch
880 Bay St. 4th Floor
Toronto, Ontario

Irving A. Singer
Smith-Singer Meteorologists, Inc.
189 Brooklyn Ave.
Massapequa, N. Y. 11758

Nozer Singpurwalla
School of Engineering
George Washington University
Washington, D. C. 20006

Bjarne Sivertsen
135 Clinton
Apt. 1U
Hemptead, New York 11550

Ralph C. Sklarew
EPA
Research Triangle Park, N. C. 277

F. Barry Smith
Meteorological Office
Met-0-14
Bracknell, Berkshire, ENGLAND

Vernon M. Smith
Box 2723
Geography Department
East Carolina University
Greenville, N. C. 27834

Ronald D. Snee
Engineering Dept.
E. I. DuPont de Nemours
Wilmington, Delaware 19898

David B. Spiegler
22 Fiske Rd.
Lexington, Mass. 02173

Robert Spirtas
Department of Biostatistics
School of Public Health
University of North Carolina
Chapel Hill, N. C. 27514

William Stasick
Dept. of Environmental Conservation
Div. of Air Resources
50 Wolf Road
Albany, New York

David J. Svendsgaard
Dept. of Biostatistics
School of Public Health
University of North Carolina
Chapel Hill, N. C. 27514

Arthur C. Stern
ESE, School of Public Health
University of North Carolina
Chapel Hill, N. C. 27514

Philip Sticksel
Battelle Columbus Laboratories
505 King Ave.
Columbus, Ohio 43201

George C. Tiao
Dept. of Statistics
University of Wisconsin 53706

George Touchton
ESE, School of Public Health
University of North Carolina
Chapel Hill, N. C. 27514

D. Bruce Turner
EPA/Div. of Meteorology
Research Triangle Park, N. C. 27711

Joseph R. Visalli
School of Civil Engineering
Purdue University
Lafayette, Indiana 47907

Raymond C. Wanta
Consulting Meteorologist
28 Hayden Lane
Bedford, Massachusetts

Lowell Wayne
Pacific Environmental Services
2932 Wilshire Blvd.
Santa Monica, California 90403

Robert Wevodau
Air Pollution Consulting Group
E. I. Dupont De Nemours
Wilmington, Delaware

J. G. Williams, Jr.
501 N. 9th St.
Room 130
Safety, Health and Welfare Bldg.
Richmond, Va. 23219

Peggy Wingard
VEPCO
P. O. Box 26666
Richmond, Virginia 23261

F. K. Wippermann
Technische Hochschule Darmstadt
Sektion Meterologie
6100 Darmstadt, GERMANY

Donald F. Worley
EPA
Research Triangle Park, N. C. 27711

Charles E. Zimmer
8160 Trabant Dr.
Cincinnati, Ohio 45242